TECHNISCHE
UNIVERSITÄT
MÜNCHEN

Basic Mathematical Tools in Imaging and Visualization

PD Dr. Tobias Lasser

lecture script
winter term 2022/23

Contents

Preface

Aim of this course. The aim of this course is two-fold. First, it aims to present and refresh selected basic mathematical tools from the areas of linear algebra, analysis, optimization and probability theory. Second, the aim is to present various applications of these tools in imaging and visualization, for example in medical imaging, image processing and computer vision.

As the range of topics of this course is quite broad, most topics are dealt with in a relatively concise manner. The overriding goal is to give you an overview over available mathematical tools, and to enable you to proceed on your own when and where necessary. Details, such as proofs, are often skipped.

Target audience of this course. The typical target audience of this course are students who just start their Master studies in some computer science related field, such as Biomedical Computing, Computational Science and Engineering, Robotics and Informatics. However, anyone who is interested is also more than welcome!

Exercises. Exercises are an integral part for understanding and practicing the concepts and methods of this course. Currently, the exercises are available separately in addition to this script, they can be downloaded from the associated Moodle course.

Feedback and comments. This course and the script for the course is an ongoing and continuing effort, and any feedback is more than welcome! Please address your feedback or comments to lasser@in.tum.de.

List of Symbols and Notations

- $\mathbb{N}$ denotes the set of natural numbers (excluding 0), i.e. $\mathbb{N} = \{1, 2, 3, \ldots\}$
- $\mathbb{N}_0$ denotes the set of natural numbers, including zero, i.e. $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$
- $\mathbb{Z}$ denotes the set of whole numbers, i.e. $\mathbb{Z} = \{0, +1, -1, +2, -2, \ldots\}$
- $\mathbb{R}$ denotes the set of real numbers
- $\mathbb{C}$ denotes the set of complex numbers
- an expression $a := b$ defines the quantity a as b, for example $X := \mathbb{N}$ defines the set X as the set of natural numbers $\mathbb{N}$
- sets are in general marked by curly braces, for example $V = \{1, 2, 3\}$, meaning V is the set containing the three numbers $1, 2, 3$
- sets drawn from bigger sets by selecting only elements with specific properties are denoted as "{big set : property}", example: $W = \{n \in \mathbb{N} : \ n \text{ odd}\}$ means W is the set of odd natural numbers
- $\times$ denotes the Cartesian product of two sets, for example $V \times V := \big\{(v_1, v_2) : \ v_1, v_2 \in V\big\}$ is the set of all pairs with elements from the set V
- mappings from one set to another are declared as follows: $f : V \to W$, $v \mapsto w$. It means the mapping is called f (or whatever is before the colon), it assigns each element of set V some element of set W, in particular mapping a specific $v \in V$ to a specific $w \in W$
- the quantor $\forall$ is a shortcut for "for all"
- the quantor $\exists$ is a shortcut for "there exists"
- the quantor $\exists!$ is a shortcut for "there exists exactly one"
- $\Rightarrow$ is the logical implication, for example $A \Rightarrow B$ means A implies B
- $\Leftrightarrow$ is the logical equivalence, for example $A \Leftrightarrow B$ means A is equivalent to B
- an expression $\langle \cdot, \cdot \rangle$ (where the dots are placeholders for two variables) typically denotes a scalar product
- an expression $\| \cdot \|$ (where the dot is a placeholder for a variable) typically denotes a norm
- an expression $\langle V \rangle$ (where V is a set) typically denotes the linear span of the set V
- $\perp$ usually means "orthogonal"
- (to be continued)

1 Linear Algebra

The concepts and methods of linear algebra are essential to almost everything related to imaging and visualization. We will start this chapter with the basic concepts of linear spaces and linear mappings, moving on to linear equation systems and various ways of solving and characterizing these linear systems. The main application example of this chapter will be in the area of computed tomography, but more application examples will be covered in the exercises.

1.1 Linear spaces

The most basic concept of linear algebra is that of a linear space. In this course we will restrict ourselves to linear spaces over the field of the real numbers $\mathbb{R}$ for simplicity. Linear spaces can also be defined over other fields (such as the complex numbers $\mathbb{C}$), but for imaging and visualization applications the real numbers are in most cases sufficient.

Definition (Linear space (over $\mathbb{R}$))**.** *Let V be a non-empty set together with the two operations "sum" $+ : V \times V \to V$ and "scalar multiplication" $\cdot : \mathbb{R} \times V \to V$, such that $a + b \in V$ and $\lambda a \in V$ for every $a, b \in V$ and $\lambda \in \mathbb{R}$. V is called a **linear space** if the following rules are fulfilled:*

(1) $a + b = b + a$ *for all* $a, b \in V$ *(commutativity)*

(2) $(a + b) + c = a + (b + c)$ *for all* $a, b, c \in V$ *(associativity)*

*(3) there exists a **zero element*** $0 \in V$ *such that* $a + 0 = a$ *for all* $a \in V$

(4) for every $a \in V$ *there exists an **inverse element*** $-a \in V$ *such that* $a + (-a) = 0$

(5) $1a = a$ *for all* $a \in V$ *(*$1 \in \mathbb{R}$ *real number)*

(6) $\lambda(\mu a) = (\lambda\mu)a$ *for all* $\lambda, \mu \in \mathbb{R}$, $a \in V$

(7) $\lambda(a + b) = \lambda a + \lambda b$ *for all* $\lambda \in \mathbb{R}$, $a, b \in V$ *(distributivity)*

(8) $(\lambda + \mu)a = \lambda a + \mu a$ *for all* $\lambda, \mu \in \mathbb{R}$, $a \in V$ *(distributivity).*

*An element $a \in V$ is called **vector**. We denote $a + (-b)$ in short as $a - b$.*

Conceptually, linear spaces describe a structure on a set V by defining the two operations, sum and scalar multiplication, which act together in a nice and expected manner.

One of the prime and very often used example of linear spaces is the typical n-dimensional space ($n \in \mathbb{N}$):

$$\mathbb{R}^n := \underbrace{\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}}_{n \text{ times}} := \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} : \; x_1, \ldots, x_n \in \mathbb{R} \right\}$$

with the operations sum

$$+ : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n, \qquad \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \mapsto \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

and scalar multiplication

$$\cdot : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n, \qquad \lambda \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

In particular, for $n = 2$ this corresponds to the intuitive two-dimensional space $\mathbb{R}^2$, and for $n = 3$ this corresponds to the intuitive three-dimensional space, which are often used in imaging and visualization.

A more general example, which actually encompasses the previous example as a special case, is the following: Let I be any non-empty set, that is $I \neq \emptyset$. Then define

$$\mathbb{R}^I := \big\{ f : \quad f : I \to \mathbb{R} \text{ mapping} \big\}.$$

$\mathbb{R}^I$ is a linear space together with the "point-wise" operations

$$\begin{aligned} (f+g)(j) &:= f(j) + g(j) \qquad \forall j \in I, \\ (\lambda f)(j) &:= \lambda f(j) \qquad \forall j \in I, \end{aligned}$$

where $f, g \in \mathbb{R}^I$, $\lambda \in \mathbb{R}$.

For particular choices of the set I, we get for example these special cases:

- Set $I = \{1, \ldots, n\}$ for $n \in \mathbb{N}$. Then $\mathbb{R}^I = \mathbb{R}^n$, as in our previous example.
- Set $I = \big\{(i,j) : \ 1 \leq i \leq n, \ 1 \leq j \leq m, \ i,j \in \mathbb{N}\big\}$ for $n, m \in \mathbb{N}$. Then we have $\mathbb{R}^I = \mathbb{R}^{m \times n}$, the linear space of $m \times n$ matrices.
- Set $I = [a,b] \subset \mathbb{R}$. Then $\mathbb{R}^{[a,b]}$ is the linear space of all real-valued functions on the compact interval $[a,b]$.

  We often look at a subset of this, namely $C([a,b])$, the set of real-valued *continuous* functions on the compact interval $[a,b]$:

  $$C([a,b]) := \big\{ f : [a,b] \to \mathbb{R} : f \text{ continuous} \big\} \subset \mathbb{R}^{[a,b]}.$$

  With the "point-wise" operations as above, $C([a,b])$ is a linear space itself as well (see exercises).

1.2 Subspaces and linear spans

Many applications involve different coordinate systems, for example in augmented reality. Before we explicitly introduce such coordinate systems, we first need to introduce subspaces as particular parts of linear spaces.

Definition (Subspace)**.** *Let V be a linear space. A nonempty set $U \subset V$ is called* ***subspace*** *of V if*

(1) $x, y \in U \quad \Longrightarrow \quad x - y \in U,$

(2) $x \in U,\ \lambda \in \mathbb{R} \quad \Longrightarrow \quad \lambda x \in U.$

U is then a linear space itself. To denote that U is a subspace of V, the short notation $U \leq V$ is used.

Checking these two requirements is sufficient. The zero element is automatically in U, as $x - x = 0 \in U$ for $x \in U$ as per (1). The same is true for the inverse element as $-1 \cdot x = -x \in U$ for $x \in U$ due to (2).

Examples:

- Let V be a linear space with zero element 0_V. Then $\{0_V\}$ is a subspace of V, in fact it is the smallest possible subspace.
- As a more concrete example we have $\mathbb{R}^2 \leq \mathbb{R}^3$, with the embedding $\mathbb{R}^2 = \left\{ \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} :\ x, y \in \mathbb{R} \right\}$.
- From our previous examples we have $C([a,b]) \leq \mathbb{R}^{[a,b]}$ (see also exercises).

Subspaces have some useful properties. Let V be a linear space. Then:

- Let U_i, $i \in I$ with I some index set, be a series of subspaces of V, then their intersection $\bigcap_{i \in I} U_i$ is also a subspace of V.
- Let $M \subset V$, then there exists a smallest subspace of V, which we denote $\langle M \rangle$, that contains M. This $\langle M \rangle$ is called the **(linear) span** of M in V. We have

  $$\langle M \rangle = \bigcap_{M \subset U \leq V} U,$$

  i.e. $\langle M \rangle$ is the intersection of all subspaces U of V that contain M, and we can also show that

  $$\langle M \rangle = \left\{ \sum_{x \in M} \lambda_x x :\ \lambda_x \in \mathbb{R},\ \text{almost all } \lambda_x = 0 \right\},$$

  ("almost all $\lambda_x = 0$" means that all except finitely many λ_x are 0), i.e. $\langle M \rangle$ is the set of all finite linear combinations of elements of M.

Using the linear span we can now define generating sets:

Definition (Generating set)**.** *Let V be a linear space. A set of vectors $(b_i)_{i \in I}$ in V is called* ***generating set*** *of V if*

$$\left\langle (b_i)_{i \in I} \right\rangle = V.$$

In other words, if the linear span of a set of vectors $(b_i)_{i \in I}$ is already the entire linear space V, then the (b_i) "generate" V.

Numerical examples: Consider $\mathbb{R}^3$ and its canonical (or standard) basis, i.e. the vectors $e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. Then $\{e_1, e_2, e_3\}$ generates $\mathbb{R}^3$, and so does $\{e_1 + e_2, e_2 + e_3, e_3\}$. The set $\{e_1, e_3\}$, however, does not generate $\mathbb{R}^3$.

1.3 Coordinate systems and bases

In order to define generalized coordinate systems, one ingredient is missing:

Definition (Linearly independent)**.** *Let V be a linear space. A finite set of vectors $a_1, \ldots, a_n \in V$ is called **linearly independent** if*

$$\lambda_1 a_1 + \ldots + \lambda_n a_n = 0 \qquad \Longrightarrow \qquad \lambda_1 = \ldots = \lambda_n = 0.$$

A set of vectors $(a_i)_{i \in I}$ in V is called linearly independent, if all finite subsets of $(a_i)_{i \in I}$ are linearly independent.

Numerical example: The canonical basis vectors of $\mathbb{R}^3$ from above, e_1, e_2, e_3, are linearly independent, as

$$\lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3 = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

implies $\lambda_1 = \lambda_2 = \lambda_3 = 0$.

Now we can define a generalized coordinate system of linear spaces, which is called "basis":

Definition (Basis)**.** *Let V be a linear space. A set of vectors $(b_i)_{i \in I}$ in V is called **basis** of V, if it is a generating set of V that is linearly independent.*

Numerical examples: Let us again consider the linear space $\mathbb{R}^3$. Then $\{e_1, e_2, e_3\}$ is a basis of $\mathbb{R}^3$, as it both generates $\mathbb{R}^3$ and is linearly independent, as shown above. The set $\{e_1, e_1 + e_2, e_1 + e_2 + e_3\}$ is also a basis of $\mathbb{R}^3$. As a counterexample, the set $\{e_1, e_2\}$ is not a basis of $\mathbb{R}^3$, as it does not generate $\mathbb{R}^3$, even though it is linearly independent. Another counterexample is the set $\{e_1, e_2, e_3, e_1 + e_2\}$, it is not a basis of $\mathbb{R}^3$ as it is not linearly independent, even though it generates $\mathbb{R}^3$.

There are some notable properties of bases, namely:

- Every linear space V has a basis.
- Let $b_1, \ldots, b_n$ be a basis of the linear space V. Then every vector $a \in V$ can be written as a **unique** linear combination $a = \sum_{i=1}^{n} \lambda_i b_i$ with $\lambda_i \in \mathbb{R}$.

  When specifying a vector, we usually put the coefficients λ_i of the linear combination of the basis vectors b_i into a "vector" of vertically stacked numbers. For example, when using the canonical basis $B_1 = \{e_1, e_2, e_3\}$ in $\mathbb{R}^3$, the vector $\left(\begin{smallmatrix} 1 \\ 2 \\ 3 \end{smallmatrix}\right)$ is the linear combination $1 \cdot e_1 + 2 \cdot e_2 + 3 \cdot e_3$.

  Caution: vectors such as $\left(\begin{smallmatrix} 1 \\ 2 \\ 3 \end{smallmatrix}\right)$ can be ambiguous, if the basis is not clearly specified. For example: let B_1 be the canonical basis from above, and let $B_2 = \{e_1 + e_2, e_2 + e_3, e_3\}$ be another basis of $\mathbb{R}^3$. Then with respect to basis B_2 the vector $\left(\begin{smallmatrix} 1 \\ 2 \\ 3 \end{smallmatrix}\right)$ means:

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}_{B_2} = 1 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + 2 \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + 3 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}_{B_1}$$

Here, we annotated the basis as a subscript to the vector to make clear which particular basis representation (or: coordinate system) we are using. Of course, the default is the canonical basis, so confusion should happen only very rarely.

- Let V be a linear space. Then we have

$$\begin{aligned} b_1, \ldots, b_n \in V \text{ basis} \quad &\iff \quad b_1, \ldots, b_n \in V \text{ minimal generating set of } V \\ &\iff \quad b_1, \ldots, b_n \in V \text{ maximal linearly independent set in } V. \end{aligned}$$

- Let V be a linear space with basis $b_1, \ldots, b_n$. Then every basis has exactly n elements. This invariant is called **dimension** of V, in short we write $\dim V := n$.

  Our frequently used example $\mathbb{R}^n$ is indeed n-dimensional, i.e. $\dim \mathbb{R}^n = n$. On the other hand, for our more general example $\mathbb{R}^I$, the dimension depends directly on the number of elements in the set I, i.e. $\dim \mathbb{R}^I = |I|$.

  For a subspace $U \leq V$ of a linear space V, we always have $\dim U \leq \dim V$ (which explains one motivation of the short-hand notation $U \leq V$).

Using bases that are well suited to our application problem is one of the major tools at our disposal. This is particularly true in image processing, where images can be represented in many fashions.

One way to represent an image is a function $f : X \to Y$, where X is the *domain*, for example 2–dimensional space $\mathbb{R}^2$ for 2–dimensional images, and Y is the *range*, which could be grey values (represented for example by $\mathbb{R}$) or RGB color values (represented for example by $\mathbb{R}^3$). Such a function f is also a vector in the linear space Y^X, and since every linear space has a basis, let's say $\{\psi_i\}$, this f can now be represented as a linear combination of these basis vectors,

$$f = \sum_i \alpha_i \psi_i,$$

where $\alpha_i \in \mathbb{R}$. The ψ_i could for example be a family of wavelets.

In a computer, a discrete representation is usually more appropriate. Instead of representing, for example, a 2-dimensional grey value image as a continuous function $f : X \to Y$ with $X = \mathbb{R}^2$ and $Y = \mathbb{R}$, we can sample X at discrete points $X' := \{1, \ldots, M\} \times \{1, \ldots, N\}$ and quantize Y into discrete values $Y' := \{1, \ldots, L\}$. Then $f' : X' \to Y'$ can be represented on a computer, for example as a $M \times N$ matrix. In fact, even f' is a vector in the linear space $\mathbb{R}^{M \times N}$ and could be represented in different fashions depending on bases of the linear space.

1.4 Linear mappings and matrices

Now that we have a structure on a set (linear spaces), as well as coordinate systems, let us look at functions that play "nice" with the structure on the set.

Definition (Linear mapping)**.** *Let V, V' be linear spaces. A mapping $f : V \to V'$ is called **linear** (or a morphism) if*

$$\begin{aligned} f(a+b) &= f(a) + f(b) \\ f(\lambda a) &= \lambda f(a) \end{aligned}$$

*for all $a, b \in V$, $\lambda \in \mathbb{R}$. A linear mapping $f : V \to V'$ is called **isomorphism** if f is bijective.*

Numerical examples:

- $0 : V \to V$, $0(v) = 0_V$ for $v \in V$ (the *zero mapping*) is linear.
- $\mathrm{id} : V \to V$, $\mathrm{id}(v) = v$ for $v \in V$ (the *identity*) is linear.
- $p_i : \mathbb{R}^n \to \mathbb{R}$, $p_i(v) = v_i$ for $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^n$ (the *projection* onto the i–th component of v) is linear.
- $t_a : V \to V$, $t_a(v) = v + a$ for $a \in V$, $v \in V$ (the *translation* by a) is **not** linear!

There is a very important theorem that allows us to reduce all finite dimensional dealings to $\mathbb{R}^n$:

Theorem. *Every ($\mathbb{R}$–)linear space V of dimension $n \in \mathbb{N}$ is isomorphic to $\mathbb{R}^n$.*

As it is instructive, here is a sketch of the proof: Let $a_1, \ldots, a_n$ be a basis of V, such that for $x \in V$ we have $x = \sum_{i=1}^n \lambda_i a_i$ with coefficients $(\lambda_i)_{i=1}^n \subset \mathbb{R}$, then

$$\varphi : V \to \mathbb{R}^n, \qquad \varphi(x) := (\lambda_i)_{i=1}^n$$

is the desired isomorphism. (We leave out the missing steps of this proof.)

Conveniently, linear mappings between finite dimensional linear spaces (which are equivalent to $\mathbb{R}^n$ with appropriate n as above) have a very useful representation:

Definition (Matrix)**.** *A vector of $\mathbb{R}^{m\times n}$ is generally written as a **matrix** $A \in \mathbb{R}^{m\times n}$,*

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}, \qquad a_{ji} \in \mathbb{R}.$$

Matrix operations. We define the usual operations on matrices briefly. Let $A, B \in \mathbb{R}^{m\times n}$ with $A = (a_{ji})$, $B = (b_{ji})$ and $\lambda \in \mathbb{R}$. Then we can define the following operations:

- Addition:
$$A + B := \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{pmatrix}$$
- Scalar multiplication:
$$\lambda A := \begin{pmatrix} \lambda a_{11} & \cdots & \lambda a_{1n} \\ \vdots & & \vdots \\ \lambda a_{m1} & \cdots & \lambda a_{mn} \end{pmatrix}$$
- Matrix multiplication with $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times r}$:
$$A \cdot B := \begin{pmatrix} \mathrm{row}_1(A)\mathrm{col}_1(B) & \cdots & \mathrm{row}_1(A)\mathrm{col}_r(B) \\ \vdots & & \vdots \\ \mathrm{row}_m(A)\mathrm{col}_1(B) & \cdots & \mathrm{row}_m(A)\mathrm{col}_r(B) \end{pmatrix}$$

or we can write it as

$$A \cdot B = \begin{pmatrix} A \cdot \mathrm{col}_1(B) & \cdots & A \cdot \mathrm{col}_r(B) \end{pmatrix}.$$

These matrix operations have many useful properties. Let $A, A_1, A_2 \in \mathbb{R}^{m\times n}$, $B, B_1, B_2 \in \mathbb{R}^{n\times r}$, $C \in \mathbb{R}^{r\times s}$. Then we have:

- $(A_1 + A_2)B = A_1B + A_2B$ and $A(B_1 + B_2) = AB_1 + AB_2$
- $\alpha(AB) = (\alpha A)B = A(\alpha B)$ for $\alpha \in \mathbb{R}$
- $A(BC) = (AB)C$
- $I_m A = AI_n = A$, where $I_n = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}$ denotes the $n \times n$ *identity matrix*
- however in general $AB \neq BA$ for $A, B \in \mathbb{R}^{n\times n}$!

Another important operation is the **transpose of a matrix.** Let $A \in \mathbb{R}^{m\times n}$ with $A = (a_{ji})$. Then

$$A^T := \begin{pmatrix} \mathrm{col}_1(A) \\ \vdots \\ \mathrm{col}_n(A) \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{mn} \end{pmatrix} \in \mathbb{R}^{n\times m}$$

is called the **transpose** of matrix A. Some useful properties are:

- $(A + B)^T = A^T + B^T$ for $A, B \in \mathbb{R}^{m\times n}$
- $(\alpha A)^T = \alpha A^T$ for $\alpha \in \mathbb{R}$, $A \in \mathbb{R}^{m\times n}$
- $(A^T)^T = A$ for $A \in \mathbb{R}^{m\times n}$
- $(AB)^T = B^T A^T$ for $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times r}$

Now that matrices are introduced, the link between linear mappings in finite dimensional linear spaces and matrices is the following theorem:

Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a linear mapping. Then by defining the matrix*

$$F = \Big(f(e_1), \ldots, f(e_n) \Big)$$

for a basis $e_1, \ldots, e_n$ of $\mathbb{R}^n$, we have

$$f(x) = Fx.$$

Conversely, to every matrix $A \in \mathbb{R}^{m\times n}$ there corresponds a linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ with $f(x) = Ax$.

In other words, every linear mapping between finite dimensional spaces can be represented by an appropriately sized matrix, and vice versa. The proof of this theorem is very simple, so we show it in full:

Proof: For $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ we have

$$\begin{aligned} f(x) = f(x_1 e_1 + \ldots + x_n e_n) &\overset{\text{linear}}{=} x_1 f(e_1) + \ldots + x_n f(e_n) \\ &\overset{\text{matr. mult.}}{=} \Big(f(e_1), \ldots, f(e_n)\Big) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = Fx. \end{aligned}$$

□

Numerical example: $f : \mathbb{R}^2 \to \mathbb{R}^3$, $f\big(\binom{x}{y}\big) = \begin{pmatrix} x \\ x+y \\ 2y \end{pmatrix}$ is a linear mapping, as we have

$$F = \Big(f(e_1), f(e_2)\Big) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 2 \end{pmatrix},$$

using the standard bases of $\mathbb{R}^2$ and $\mathbb{R}^3$. We also have

$$f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ x+y \\ 2y \end{pmatrix}.$$

1.5 Linear systems and matrices

Very many applications in imaging and visualization reduce to solving linear systems after mathematical modeling. This is why we will spend considerable time on ways to solve them.

Definition (Linear system)**.** *A **linear equation** has the form*

$$a_1 x_1 + \ldots a_n x_n = b,$$

where $x_1, \ldots, x_n \in \mathbb{R}$ are the unknowns/variables and $a_1, \ldots, a_n, b \in \mathbb{R}$ are constants.

*A **linear system** is a finite set of m linear equations with variables $x_1, \ldots, x_n \in \mathbb{R}$:*

$$\begin{array}{rcl} a_{11}x_1 + \ldots + a_{1n}x_n & = & b_1 \\ \vdots \qquad\qquad \vdots & & \vdots \\ a_{m1}x_1 + \ldots + a_{mn}x_n & = & b_m \end{array}$$

or in matrix notation

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

In short

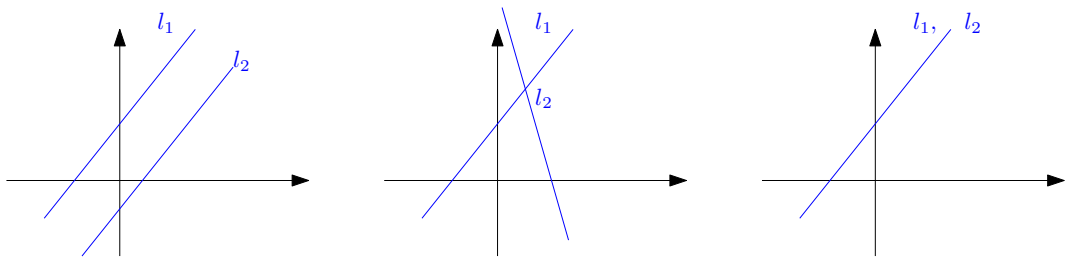
$$Ax = b,$$

with $A = (a_{ji})$ and $b = (b_j)$.

Property: A linear system always has either no solution, exactly one solution or infinitely many solutions.

This property is illustrated by this simple example of a pair of straight lines in $\mathbb{R}^2$, which can be modeled as a linear equation each:

$$\begin{array}{lll} (l_1) & a_1x + b_1y = c_1 & (a_1, b_1) \neq (0,0) \\ (l_2) & a_2x + b_2y = c_2 & (a_2, b_2) \neq (0,0) \end{array}$$



no solution | exactly one solution | infinite many solutions

The case of no solution corresponds here to parallel lines, the case of exactly one solution to the lines intersecting, and the case of infinitely many solutions to the case of completely overlapping lines.

One of the easiest ways to solve linear systems for the unknown variables is to know the inverse of the matrix.

Definition (Inverse matrix)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix. If there exists $B \in \mathbb{R}^{n\times n}$ with*

$$AB = BA = I_n,$$

*(where I_n denotes the $n\times n$ identity matrix), then A is called **invertible**, and we call $B := A^{-1}$ the **inverse** of A.*

For small matrices there are explicit formulas to compute the inverse (for bigger matrices, there are unfortunately no explicit formulas). Here is the example of a 2×2 matrix: Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. If $ad - bc \neq 0$, then A is invertible with

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Numerical example: set $A = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$, then $A^{-1} = \frac{1}{-2}\left(\begin{smallmatrix} 4 & -2 \\ -3 & 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{smallmatrix}\right)$. We can check that this is indeed the inverse:

$$AA^{-1} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = A^{-1}A.$$

There are some useful properties of invertible matrices:

- If the inverse of a matrix exists, it is unique.

- $A, B \in \mathbb{R}^{n\times n}$ invertible $\iff$ AB invertible and $(AB)^{-1} = B^{-1}A^{-1}$.
- $A \in \mathbb{R}^{n\times n}$ invertible $\iff$ $Ax = 0$ has only the trivial solution $x = 0$
  $\iff$ $Ax = b$ has exactly one solution for every $b \in \mathbb{R}^n$, $x = A^{-1}b$.

The latter in particular is one of the easiest ways of solving a linear system $Ax = b$. If A is invertible and the inverse A^{-1} is known, just compute $x = A^{-1}b$.

Unfortunately, except in very simple matrices, it is not easy to determine if a matrix is invertible. There are, however, several tools that can help. One of them is the determinant:

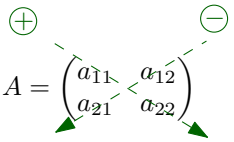
Definition (Determinant)**.** *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix, $A = (a_{ji})$. The **determinant** of A is defined as*

$$\det(A) := \sum_{\sigma\in S_n} sgn(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)},$$

where S_n is the set of all permutations of $\{1, \dots, n\}$.

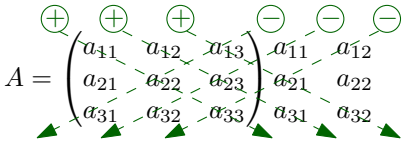
This is the so-called Leibniz formula for determinants, which is one of the shortest ways to introduce the determinant. Unfortunately, it is really impractical to use to actually compute determinants. As before, for small matrices, there are easy-to-use schemes to compute the determinant of a matrix $A \in \mathbb{R}^{n\times n}$:

- $n = 2$:



  $\implies \det(A) = a_{11}a_{22} - a_{12}a_{21}.$

- $n = 3$:



$$\implies \det(A) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}.$$

In case your matrix is not that small, there are some useful properties that can help your computations. Let $A \in \mathbb{R}^{n\times n}$.

- If $A = (a_{ji})$ is a triangular $\left(\begin{smallmatrix}\ast & \ast\\ & \ast\end{smallmatrix}\right)$, $\left(\begin{smallmatrix}\ast & \\ \ast & \ast\end{smallmatrix}\right)$ or diagonal $\left(\begin{smallmatrix}\ast & \\ & \ast\end{smallmatrix}\right)$ matrix, then

$$\det(A) = a_{11}a_{22}\cdots a_{nn}.$$

- $\det(A) = \det(A^T)$.
- $\det(\lambda A) = \lambda^n \det(A)$ for $\lambda \in \mathbb{R}$.

- $A, B \in \mathbb{R}^{n\times n}$, then $\det(AB) = \det(A)\det(B)$.
- A invertible $\iff$ $\det(A) \neq 0$.
- A invertible, then $\det(A^{-1}) = \frac{1}{\det(A)}$.

In particular, we see that a matrix with non-zero determinant is invertible.

We now introduce more properties of matrices:

Definition (Image)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a mapping. The **image** of f is defined as*

$$Im(f) := \big\{b \in \mathbb{R}^m : \ \exists x \in \mathbb{R}^n : \ f(x) = b\big\}.$$

*Correspondingly, let $A \in \mathbb{R}^{m\times n}$, then the **image** of A is*

$$Im(A) = \big\{b \in \mathbb{R}^m : \ \exists x \in \mathbb{R}^n : \ Ax = b\big\}.$$

The image of A is generated by linear combinations of the columns of A,

$$b = Ax = \big(col_1(A) \quad \cdots \quad col_n(A)\big)\, x = col_1(A)x_1 + \cdots + col_n(A)x_n.$$

*Hence we also call $Im(A)$ the **column space** of A, the subspace of $\mathbb{R}^m$ spanned by $\{col_i(A)\}_{i=1}^n$. Correspondingly, the **row space** of A is the subspace of $\mathbb{R}^n$ spanned by $\{row_j(A)\}_{j=1}^m$.*

This allows us to formulate some properties. Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$.

- A linear system $Ax = b$ has a solution $\iff$ b is contained in column space of A, i.e., $\mathrm{Im}(A)$.
- The column and row space of A always have the same dimension.

This latter feature is remarkable, and allows us to define the rank of a matrix:

Definition (Rank)**.** *Let $A \in \mathbb{R}^{m\times n}$. The dimension of the column and row space of A is called **rank** of A, rank(A).*

Closely related to the rank of a matrix is its kernel:

Definition (Kernel)**.** *Let $A \in \mathbb{R}^{m\times n}$. The set*

$$\ker(A) := \big\{x \in \mathbb{R}^n : \ Ax = 0\big\}$$

*is a subspace of $\mathbb{R}^n$ and is called **null space** of A or **kernel** of A.*

The rank and the kernel of a matrix are important properties of a matrix. Let $A \in \mathbb{R}^{m\times n}$. Then:

- $\mathrm{rank}(A) = 0 \iff A = 0$.
- $\mathrm{rank}(A) \leq \min(m, n)$.
- $n - \mathrm{rank}(A) = \dim\ker(A)$.

- If $m = n$, i.e. the matrix is square, then:
  A invertible $\iff$ $\operatorname{rank}(A) = n$ ("A has full rank")
  $\iff$ $\dim \ker(A) = 0$ ($\iff$ $\ker(A) = \{0\}$).

Again, the latter feature is important for the solution of linear systems. If the rank is full or the kernel just contains the zero vector, then the linear system with matrix A has exactly one solution.

Numerical example: We can use the properties introduced earlier (determinant, rank, kernel) to determine whether a matrix is invertible.

- $A = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$, then $\det(A) = 4 - 6 = -2 \implies$ A invertible, and $\operatorname{rank}(A) = 2$.
- $A = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right)$ has linearly dependent columns, e.g. $2\left(\begin{smallmatrix} 1 \\ 2 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 2 \\ 4 \end{smallmatrix}\right) \implies \operatorname{rank}(A) = 1 \implies$ A not invertible. In addition: $\det(A) = 4 - 4 = 0$, which also implies non-invertibility.

1.6 Lengths, distances and angles

Now that we have introduced vectors in linear spaces, we want to do something with these vectors. Everyday tasks like measuring lengths, distances or angles are very useful to have for vectors in linear spaces as well. In this section we will introduce the standard approaches on how to do this.

First we introduce the length of a vector:

Definition (Norm)**.** *Let V be a linear space (over $\mathbb{R}$). A mapping*

$$\|\cdot\| : V \to \mathbb{R}$$

is called ***norm*** *on V if it fulfills*

(1) $\|x\| \geq 0$ *and* $\|x\| = 0 \;\Leftrightarrow\; x = 0$

(2) $\|\lambda x\| = |\lambda| \, \|x\|$

(3) $\|x + y\| \leq \|x\| + \|y\|$

for $x, y \in V$ and $\lambda \in \mathbb{R}$. The first property is called positive definiteness, the second one homogeneity and the third one is called the triangle inequality.

The main examples of a norm are the so-called p**-norms** on the linear spaces $\mathbb{R}^n$ ($n \in \mathbb{N}$): Let $x = (x_i) \in \mathbb{R}^n$ and $p \geq 1$, then the p-norm of x is defined as

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

For $p = 2$ this corresponds to real-world length measurements, we also call the 2-norm the **Euclidean norm**. The length of the two-dimensional vector $a = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right) \in \mathbb{R}^2$, for example, computes as $\|a\|_2 = \sqrt{1^2 + 0^2} = 1$, while $b = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right) \in \mathbb{R}^2$ has length $\|b\|_2 = \sqrt{1^2 + 1^2} = \sqrt{2}$. Of course there are also many other norms on many other linear spaces.

Using the length, i.e. the norm of a vector, we can also introduce a notion of distance between two vectors by computing the length of the difference between the two vectors:

Definition (Distance)**.** *Let V be a linear space with a norm $\|\cdot\|$. Then for $x, y \in V$ the* ***distance*** *of x and y is defined as*

$$d(x, y) := \|y - x\|.$$

Following the previous example in $\mathbb{R}^2$ with the 2-norm, we can compute the distance between vectors $c = \binom{3}{2} \in \mathbb{R}^2$ and $d = \binom{1}{3} \in \mathbb{R}^2$ as $d(c, d) = \| \binom{1-3}{3-2} \|_2 = \sqrt{2^2 + 1^2} = \sqrt{5}$. As the 2-norm is called Euclidean norm, the corresponding distance is called **Euclidean distance**.

Distances defined like this are also a *metric*, which is a central concept of the next chapter (Analysis), but more on that later.

Finally, it is also very useful to have a notion of angle between two vectors. While angle is an intuitive concept in two dimensions ($\mathbb{R}^2$), it is a bit more involved to extend this to higher dimensions or arbitrary linear spaces. The main mathematical tool to do this is the so-called "scalar product":

Definition (Scalar product)**.** *Let V be a linear space (over $\mathbb{R}$). A mapping*

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

is called ***scalar product*** *(or* ***inner product*** *or* ***dot product****) on V if it fulfills*

(1) $\langle x + x', y \rangle = \langle x, y \rangle + \langle x', y \rangle$ *and* $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$

(2) $\langle x, y \rangle = \langle y, x \rangle$

(3) $\langle x, x \rangle \geq 0$ *and* $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

for all $x, x', y \in V$ and $\lambda \in \mathbb{R}$. The first property is called linearity, the second symmetry and the third one positive definiteness.

The standard example of a scalar product is the *standard inner product* on $\mathbb{R}^n$, which is defined for vectors $x = (x_i), y = (y_i) \in \mathbb{R}^n$ as follows:

$$\langle x, y \rangle := \sum_{i=1}^{n} x_i y_i.$$

An equivalent short notation is $\langle x, y \rangle = x^T y$.

Numerical example: using the same vectors $c = \binom{3}{2}$ and $d = \binom{1}{3}$ in $\mathbb{R}^2$ as before: $\langle c, d \rangle = 3 \cdot 1 + 2 \cdot 3 = 9$.

Other scalar products can be defined on $\mathbb{R}^n$ (more on that later), or on other linear spaces. One such example is the linear space $C([a, b])$ from above, here

$$\langle f, g \rangle := \int_a^b f(t) g(t)\, dt$$

for $f, g \in C([a, b])$, defines a scalar product. There are some caveats with linear spaces of functions, the scalar product via an integral of the product of functions does not always work. For example, in general it does **not** work for $\mathbb{R}^{[a,b]}$ due to undefined integrals.

If a linear space V has a scalar product $\langle \cdot, \cdot \rangle$, you can automatically define a norm on V,

$$\|x\| := \sqrt{\langle x, x \rangle}$$

for $x \in V$. We call this norm the *norm induced by the scalar product.*

Recalling the p-norms on $\mathbb{R}^n$ from before, for $p = 2$ the norm is obviously induced by the standard inner product. For $p \neq 2$ the p-norm is not induced by any scalar product (as can be shown via the parallelogram law).

Going back to a general linear space V with scalar product $\langle \cdot, \cdot \rangle$: An important tool using the scalar product is the **Cauchy-Schwarz Inequality**:

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \ \langle y, y \rangle \qquad \forall x, y \in V.$$

We usually do not cover proofs in this course, however the proof of this inequality is quite instructive and hence posed as an exercise (see exercise sheets).

We use the Cauchy-Schwarz inequality now in order to introduce the notion of angle between vectors:

Definition (Angle)**.** *Let V be a linear space with a scalar product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $x, y \in V$, where $x, y \neq 0$, then we have*

$$\langle x, y \rangle = \|x\| \ \|y\| \cos \varphi,$$

*where φ is the **angle** between x and y.*

In particular, $\langle x, y \rangle = 0$ implies $\varphi = \frac{\pi}{2} = 90°$, i.e. x, y are orthogonal or in short notation: $x \perp y$.

This angle is a well-defined quantity, as is proven in the following: For $x, y \in V$ with $x, y \neq 0$ we have due to the Cauchy-Schwarz inequality that $\langle x, y \rangle^2 \leq \|x\|^2 \ \|y\|^2$, and thus by applying the square root $-\|x\| \ \|y\| \leq \langle x, y \rangle \leq \|x\| \ \|y\|$, which yields

$$-1 \leq \frac{\langle x, y \rangle}{\|x\| \ \|y\|} \leq 1.$$

$\cos : [0, \pi] \to \mathbb{R}$ is strongly monotonously decreasing from 1 to -1. Thus there exists exactly one $(!\exists)$ $\varphi \in [0, \pi]$ with

$$\varphi = \cos^{-1} \left(\frac{\langle x, y \rangle}{\|x\| \ \|y\|} \right).$$

Numerical examples: For our previous example vectors $c, d \in \mathbb{R}^2$ with $c = \left(\begin{smallmatrix} 3 \\ 2 \end{smallmatrix}\right)$, $d = \left(\begin{smallmatrix} 1 \\ 3 \end{smallmatrix}\right)$, we have $\frac{\langle c, d \rangle}{\|c\| \ \|d\|} = \frac{9}{\sqrt{130}}$, and thus $\varphi = \cos^{-1} \left(\frac{9}{\sqrt{130}} \right) = 0.661\ldots \approx 37.9°$. (You are welcome to double-check this by a drawing on paper and actually measuring the angle between the two vectors.)

As this notion of angle is now defined in general for any linear space with scalar product and induced norm, we can also (for fun) apply it for example to the linear space $C([0, 1])$ with the previously introduced scalar product $\langle f, g \rangle := \int_0^1 f(t) g(t) \ dt$ for $f, g \in C([0, 1])$. The angle between the two example functions $f(t) = t$ and $g(t) = t^2$ can then be computed as $\varphi \approx 14.5°$ (whatever that actually means).

1.7 Solving linear systems: Computed Tomography

As one of the prime application examples in this course we consider Transmission X–ray Computed Tomography (X-ray CT). Here the goal is to recover a function $f : V \to \mathbb{R}$ that maps a volume of interest $V \subset \mathbb{R}^n$ to some real–valued property (in X-ray CT that is the X–ray attenuation coefficient of the object that is measured).

One typical approach to this problem is the so-called "series expansion" approach outlined below:

- **Step 1:** Discretize f.

  Choose a finite set of "basis functions" $b_i : V \to \mathbb{R}$, $i \in I$, such that f can be approximated by a linear combination $\hat{f}$ of "basis functions" b_i. (The b_i can be a real basis, but that is not necessary.) That is:

  $$\hat{f}(\cdot) = \sum_{i \in I} x_i b_i(\cdot), \qquad \{x_i\}_{i \in I} \subset \mathbb{R},$$

  such that $\left\| f - \hat{f} \right\| < \varepsilon$, where $\varepsilon > 0$ small.

  The vector $x = (x_i)_{i \in I}$ is the discretized version of the quantity to be reconstructed, f. If the b_i are linearly independent, the mapping $\hat{f} \mapsto x$ is bijective.

  Standard choice for 2D case: $k \times k$ pixel grid, that is $I = \{1, \ldots, k^2\}$ and

  $$b_i(c_1, c_2) := \begin{cases} 1 & \text{if } (c_1, c_2) \text{ is inside } i\text{–th pixel} \\ 0 & \text{else.} \end{cases}$$

- **Step 2:** Decide on a measurement model $\mathcal{M}_j$.

  Let $m = (m_j)_{j \in J} \subset \mathbb{R}$ denote the finite set of measured values of the detector. We need a model of the measurement process

  $$\mathcal{M}_j : \big(f : V \to \mathbb{R}\big) \longrightarrow \mathbb{R},$$

  which describes how the property f could generate our measured signals m, such that $\mathcal{M}_j f = m_j$ for all $j \in J$. Assuming $\mathcal{M}_j$ is linear, we have

  $$\mathcal{M}_j f \approx \mathcal{M}_j \hat{f} = \mathcal{M}_j \left(\sum_{i \in I} x_i b_i \right) = \sum_{i \in I} x_i \, \mathcal{M}_j b_i.$$

  In our example of X-ray CT a simple model is the X-ray transform: $\mathcal{M}_j f = \int_{L_j} f(x) \, dx$, which is just a line integral along the line L_j the X-ray took from source to detector.

  This allows us to form a system equation

  $$Ax = m$$

  using the system matrix $A = (a_{ji})$ with $a_{ji} = \mathcal{M}_j b_i$.

- **Step 3:** Solve the system equation $Ax = m$.

  Compute a solution $\hat{x}$ (or approximation to solution) such that $\hat{f}^* = \sum_{i \in I} \hat{x}_i b_i$ is the desired reconstruction of f.

Each of the three steps has many choices for implementation. Of particular interest for us right now is step 3, which involves solving a linear system.

In the following we present a method for solving such a linear system (Kaczmarz's method), which requires one more mathematical tool:

Orthogonal projections. To project a vector $x \in \mathbb{R}^n$ orthogonally onto a unit vector $u \in \mathbb{R}^n$ (a unit vector has length 1, i.e. $\|u\| = 1$) we use the projection operator

$$P_u : \mathbb{R}^n \to \mathbb{R}^n, \qquad P_u(x) = \langle x, u \rangle u$$

P_u has the following useful properties:

- We have $P_u(u) = \langle u, u \rangle u = u$.
- For $x = x_\parallel + x_\perp$, where
  - $x_\parallel$ is the part of x which is parallel to u and
  - $x_\perp$ is the part that is orthogonal to u

  we have

$$P_u(x) = P_u(x_\parallel + x_\perp) = \langle x_\parallel, u \rangle u + \langle x_\perp, u \rangle u = \|x_\parallel\| u + 0u = x_\parallel.$$

Now we can introduce a method for solving linear systems.

Kaczmarz's method. (also called ART — Algebraic Reconstruction Technique)
Let

$$Ax = m$$

be a linear system with $A \in \mathbb{R}^{l \times n}$, $A \neq 0$, $m = (m_j) \in \mathbb{R}^l$ and unknowns $x \in \mathbb{R}^n$.

Denote $a_j := \mathrm{row}_j(A)$, then each row $a_j x = m_j$ of the linear system forms an affine hyperplane in $\mathbb{R}^n$,

$$H_j := \big\{ x \in \mathbb{R}^n : \ \langle a_j, x \rangle = m_j \big\}.$$

(A hyperplane of $\mathbb{R}^n$ is a subspace of dimension $n - 1$ and divides $\mathbb{R}^n$ in half.) We have

$$x = \bigcap_{j=1}^{l} H_j,$$

that means the solution of the linear system is the intersection of all affine hyperplanes H_j (think back to our pair of lines example earlier).

Kaczmarz's method is now to compute x by successively projecting onto the affine hyperplanes H_j using the projection operator introduced above, until we converge to the solution. If P_{H_j}

denotes the projection onto H_j, then one iteration of Kaczmarz's method can be written as $\Phi = P_{H_l} \circ \ldots \circ P_{H_1}$. In full, the method reads

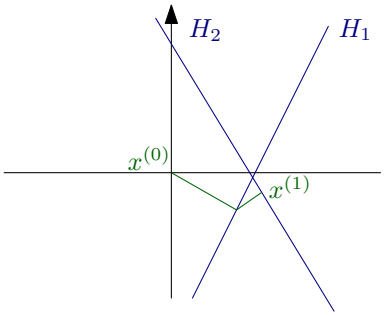
$$x^{(0)} = 0 \qquad \text{(or any other starting value)}$$

iterate over $k = 0, \ldots, \hat{k}$

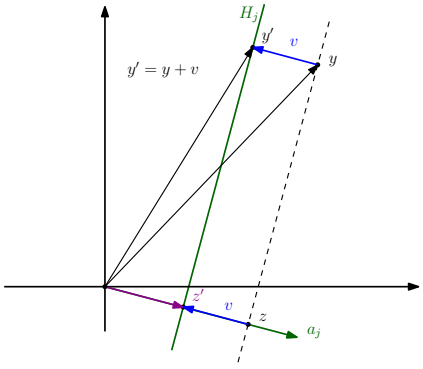
$$x^{(k+1)} = \Phi \; x^{(k)}$$

In case $Ax = m$ has a solution, this method will indeed converge to a solution.

Here is an example illustration of one iteration of the Kaczmarz method in $\mathbb{R}^2$ with two affine hyperplanes H_1, H_2:



In order to calculate P_{H_j} we denote $y := P_{H_{j-1}} \circ P_{H_{j-2}} \circ \ldots \circ P_{H_1} x^{(k)}$ and $y' := P_{H_j} y$.



In order to obtain $y' = P_{H_j} y$ we try to find a correction v such that $y' = y + v$. As y' is the projection of y onto H_j we can expect that v will be parallel to a_j, because a_j is the normal vector of (and thus orthogonal to) H_j.

At first, we note that v can be obtained as

$$v = z' - z,$$

where z is the orthogonal projection of y onto $a_j / \|a_j\|_2$ and z' is the orthogonal projection of z onto H_j. Thus, we can compute y' via the following three steps:

1. Compute z: The orthogonal projection of y onto $a_j / \|a_j\|_2$ can be obtained by employing the definition of an orthogonal projection, i.e.

$$z = \left\langle y, \frac{a_j}{\|a_j\|_2} \right\rangle \frac{a_j}{\|a_j\|_2}.$$

2. Compute z': z' has to be parallel to $a_j / \|a_j\|_2$ and has to be an element of H_j. Thus, it must be a scaled version of $a_j / \|a_j\|_2$. Choosing $z' = m_j a_j / \|a_j\|_2^2$ yields the desired vector as the following calculation shows:

$$\langle z', a_j \rangle - m_j = \left\langle m_j \frac{a_j}{\|a_j\|_2^2}, a_j \right\rangle - m_j = m_j \frac{\langle a_j, a_j \rangle}{\|a_j\|_2^2} - m_j = m_j - m_j = 0.$$

3. Compute y': We just have to combine the observations made above:

$$y' = y + v = y + z' - z = y + \frac{m_j - \langle y, a_j \rangle}{\|a_j\|_2^2} a_j.$$

1.8 Orthogonality

Orthogonality is an important concept, as we have already seen in Kaczmarz's method. In this section we will add a few more mathematical tools related to orthogonality.

In order to define what *orthogonal* means, we first require the scalar product from the previous section. One notable example in $\mathbb{R}^n$ ($n \in \mathbb{N}$) is the *standard scalar product*,

$$\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \qquad \langle x, y \rangle := \sum_{i=1}^{n} x_i y_i$$

for $x = (x_i), y = (y_i) \in \mathbb{R}^n$.

As an aside, we also recall that it induces the Euclidean norm $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}$, $\|x\| := \sqrt{\langle x, x \rangle}$, which induces a distance measure (metric) $d(x, y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $d(x, y) := \|y - x\|$. The scalar product also induces a notion of angle

$$\varphi = \cos^{-1}\left(\frac{\langle x, y \rangle}{\|x\| \ \|y\|} \right).$$

We will restrict ourselves in the following to finite dimensional linear spaces, hence considering $\mathbb{R}^n$ for arbitrary $n \in \mathbb{N}$ is sufficient.

Definition (Orthogonal)**.** *Let $x, y \in \mathbb{R}^n$. x and y are called **orthogonal** (or: $x \perp y$) if $\langle x, y \rangle = 0$.*

Using the angle φ, $\langle x, y \rangle$ describes an angle of 90° between x and y. Orthogonality can be particularly useful in bases:

Definition (Orthonormal basis (ONB))**.** *A basis $(b_i)_{i=1}^n$ of $\mathbb{R}^n$ is called **orthonormal**, if all basis vectors are pairwise orthogonal and have norm 1, in short:*

$$\langle b_i, b_j \rangle = \delta_{ij} := \begin{cases} 1 & i = j \\ 0 & else \end{cases}$$

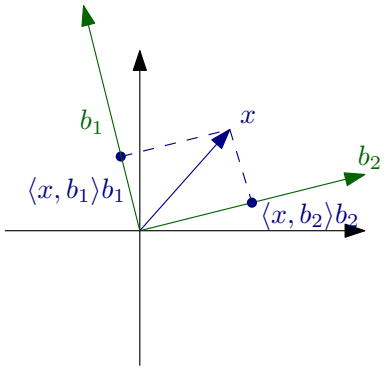
for all $i, j \in \{1, \ldots, n\}$.

Numerical example: Let's take $\mathbb{R}^3$ with the standard basis $e_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}$, $e_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}$, $e_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}$. We have $\|e_1\| = \|e_2\| = \|e_3\| = 1$ and $\langle e_1, e_2\rangle = \langle e_1, e_3\rangle = \langle e_2, e_3\rangle = 0$. Thus $\{e_1, e_2, e_3\}$ is an orthonormal basis (ONB) of $\mathbb{R}^3$.

Property: Let $\{b_1, \ldots, b_n\}$ be ONB of $\mathbb{R}^n$. Then we have for $x \in \mathbb{R}^n$:

$$x = \langle x, b_1\rangle b_1 + \ldots + \langle x, b_n\rangle b_n.$$

That is: the coefficient of each basis vector is the projection onto that basis vector (recall the projection operation from the previous section).



We show the proof because of its simplicity:

Proof: Let $x = \sum_{i=1}^n \alpha_i b_i$. Then for $j = 1, \ldots, n$

$$\langle x, b_j\rangle = \left\langle \sum_{i=1}^n \alpha_i b_i, b_j \right\rangle = \sum_{i=1}^n \alpha_i \langle b_i, b_j\rangle = \sum_{i=1}^n \alpha_i \delta_{ij} = \alpha_j.$$

□

Another helpful tool related to orthogonality are orthogonal matrices:

Definition (Orthogonal matrix)**.** *A square matrix $A \in \mathbb{R}^{n\times n}$ is called **orthogonal** if $A^{-1} = A^T$, or alternatively $A^TA = AA^T = I_n$.*

Orthogonal matrices have some neat properties. Consider square matrices $A, B \in \mathbb{R}^{n\times n}$.

- A orthogonal $\implies$ rows of A are orthonormal in $\mathbb{R}^n$
  $\implies$ columns of A are orthonormal in $\mathbb{R}^n$
- A, B orthogonal $\implies$ AB orthogonal
- A orthogonal $\implies$ $|\det(A)| = 1$.
- A orthogonal $\iff$ $\|Ax\| = \|x\|$ for all $x \in \mathbb{R}^n$ (length preserving)

Example: Rotation in $\mathbb{R}^2$ counterclockwise by angle θ,

$$A = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix},$$

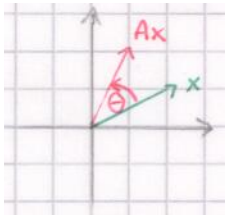
is orthogonal, as

$$A^T A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = AA^T,$$

using $(\cos\theta)^2 + (\sin\theta)^2 = 1$. We also have $\det(A) = (\cos\theta)^2 + (\sin\theta)^2 = 1$.

Numerical example: Rotate $x = \binom{2}{1}$ by $\theta \approx 36.87°$:

$$Ax = \begin{pmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$



1.9 Special matrices and solving linear systems

As it is so important for very many applications, let us consider again a linear system

$$Ax = b$$

with $A \in \mathbb{R}^{n\times n}$ square, $b \in \mathbb{R}^n$ and unknowns $x \in \mathbb{R}^n$. When is this system solvable? We already know some criteria:

- $b \in \mathrm{Im}(A) \quad \Longrightarrow \quad$ at least one solution exists
- $\mathrm{rank}(A) = n \quad \Longrightarrow \quad$ exactly one solution exists
- $\det(A) \neq 0 \quad \Longrightarrow \quad$ exactly one solution exists
- $\ker(A) = \{0\} \quad \Longrightarrow \quad$ exactly one solution exists

For the last three cases, A is invertible and $x = A^{-1}b$ is the solution.

How do we actually compute the inverse for that?

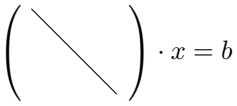
- use explicit formula for $n = 2, 3, 4$
- compute directly via Gauss-Jordan elimination
- compute indirectly via decompositions of A (e.g. LU or QR)

Calculating the inverse directly, however, is very likely to have numerical issues, so it is not advised to do that.

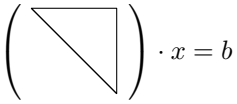
Numerical example: $\begin{pmatrix} 1.9998 & 0.9999 \\ 3.9994 & 2.0009 \end{pmatrix}$ is invertible, but $\begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$ is not. Numerical errors (e.g. rounding errors or floating point limitations) can destroy the result.

How do we avoid computing the inverse then? If A has a "nice" shape, the system can be solved easily:

- A diagonal $\implies$ read off solution

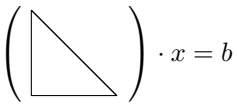


- A upper triangular $\implies$ use backward substitution (n^2 FLOPs)



U for 'upper'

- A lower triangular $\implies$ use forward substitution (n^2 FLOPs)



L for 'lower'

But then, our matrix A is rarely in such a "nice" shape. So how do we actually solve a linear system $Ax = b$ with $A \in \mathbb{R}^{n \times n}$ in the end? There are three main approaches:

- direct solution via **inverse** A^{-1} (however A might not be invertible, and computing the inverse has numerical issues)
- solution via **decomposition** of A into "nice" shapes (for example $A = QR$, a decomposition into orthogonal and upper triangular matrix, this is numerically stable, but again only works for invertible A)
- **iterative methods** for step by step approximation of the solution (very suitable for huge n, also works in case there is no solution $\rightarrow$ least squares problems)

So if A can be decomposed into "easy" factors, a solution can easily be calculated. In the following we shortly present some of these matrix decompositions:

Theorem (Cholesky decomposition)**.** *Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive definite (i.e. $A = A^T$ and $\langle x, Ax \rangle > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$). Then there exists a unique lower triangular matrix $L \in \mathbb{R}^{n \times n}$ with strictly positive diagonal entries and*

$$A = LL^T.$$

- Using the Cholesky decomposition, we solve $LL^T x = b$ instead of $Ax = b$ by
  1. solving $Lz = b$ for z using forward substitution
  2. solving $L^T x = z$ for x using backward substitution
- cost: $\sim \frac{n^3}{3}$ FLOPs, can be computed in-place

- might have numerical stability issues
- If you drop the requirements on A to symmetric, positive semi–definite, then the uniqueness and the strictly positive diagonal entries drop out of the theorem as well.

Theorem (LU decomposition)**.** *Let $A \in \mathbb{R}^{n\times n}$ be invertible and let A^T be diagonally dominant ($|a_{ii}| \geq \sum_{i\neq j} |a_{ji}|$ for all $i = 1, \ldots, n$). Then there exists a decomposition*

$$A = LU,$$

where $L \in \mathbb{R}^{n\times n}$ lower triangular and $U \in \mathbb{R}^{n\times n}$ upper triangular. If either L or U has only ones on the diagonal, the decomposition is unique.

- Using the LU decomposition, we solve $LUx = b$ instead of $Ax = b$ by
  1. solving $Lz = b$ for z using forward substitution
  2. solving $Ux = z$ for x using backward substitution
- cost: $\sim \frac{2n^3}{3}$ FLOPs, can be computed in-place
- might have numerical issues, such as division by zero or very small numbers

Theorem (LUP decomposition)**.** *Let $A \in \mathbb{R}^{n\times n}$ be invertible. Then there exists a unique decomposition*

$$PA = LU,$$

where $P \in \mathbb{R}^{n\times n}$ is a permutation matrix, $L \in \mathbb{R}^{n\times n}$ lower triangular with unit diagonal and $U \in \mathbb{R}^{n\times n}$ upper triangular.

- Using the LUP decomposition, we solve $(PAx =)LUx = Pb$ instead of $Ax = b$ by
  1. solving $Lz = Pb$ for z using forward substitution
  2. solving $Ux = z$ for x using backward substitution
- cost: $\sim \frac{2n^3}{3}$ FLOPs, can be computed in-place using an additional vector $\in \mathbb{R}^n$ to store the permutations
- this method is numerically stable

Theorem (QR decomposition)**.** *Let $A \in \mathbb{R}^{n\times n}$ be invertible. Then there exists a unique orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ and a unique upper triangular matrix $R \in \mathbb{R}^{n\times n}$ with positive diagonal entries and*

$$A = QR.$$

- Using the QR decomposition, we solve $Rx = Q^T b$ instead of $Ax = b$ by
  1. solving $Rx = Q^T b$ for x using backward substitution
- cost: $\sim \frac{2n^3}{3}$ FLOPs
- this is also numerically stable

- If $A \in \mathbb{R}^{m\times n}$, $m > n$, then there exists $Q \in \mathbb{R}^{m\times m}$ orthogonal and $R \in \mathbb{R}^{m\times n}$ upper triangular with
$$A = QR = Q\begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$
with $R_1 \in \mathbb{R}^{n\times n}$ upper triangular. Alternatively
$$A = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}\begin{pmatrix} R_1 \\ 0 \end{pmatrix} = Q_1 R_1$$
with $Q_1 \in \mathbb{R}^{m\times n}$, $Q_2 \in \mathbb{R}^{m\times(m-n)}$, Q_1, Q_2 have orthonormal columns. Q_1R_1 is called **thin** or **reduced** QR factorization.

  If $\operatorname{rank}(A) = n$ and R_1 has positive diagonal entries, then R_1 and Q_1 are unique.

Other decompositions exist (for example the SVD, see next section). Sometimes even these decomposition methods are not suitable (because of performance, or sparse matrices, or numerical issues). If that is the case, then there remains the third approach to solving linear system, using iterative methods (see for eaxmple the Kaczmarz/ART method from before, or the CG method later).

1.10 Singular value decomposition

Theorem (Singular value decomposition)**.** *Let $A \in \mathbb{R}^{m\times n}$. Then there exists a decomposition of A, the so-called **singular value decomposition** (SVD),*
$$A = U\Sigma V^T.$$
*For $m \geq n$, we have $U = (u_1, \ldots, u_n) \in \mathbb{R}^{m\times n}$ has orthogonal columns (i.e. $U^TU = I_n$, but in general $UU^T \neq I_m$), $V = (v_1, \ldots, v_n) \in \mathbb{R}^{n\times n}$ is orthogonal and $\Sigma = diag(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$ is diagonal with non–negative entries sorted in non–increasing order, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. The σ_i are called **singular values**. U is also length preserving, $\|Ux\| = \|x\|$ for all $x \in \mathbb{R}^n$.*

For $m < n$, we have $U = (u_1, \ldots, u_m) \in \mathbb{R}^{m\times m}$ orthogonal, $\Sigma = diag(\sigma_1, \ldots, \sigma_m) \in \mathbb{R}^{m\times m}$ with $\sigma_1 \geq \ldots \sigma_m \geq 0$, and $V = (v_1, \ldots, v_m) \in \mathbb{R}^{n\times m}$ with orthogonal columns.

Property: Let $A \in \mathbb{R}^{n\times n}$ square with SVD $A = U\Sigma V^T$. Then $A^TA = V\Sigma U^TU\Sigma V^T = V\Sigma^2V^T$ and thus $V^TA^TAV = \Sigma^2$. Thus: the singular values of A are the square roots of the eigenvalues of A^TA (which is symmetric!), see the next sections.

Applications:

- Solving linear equations $Ax = b$ for $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \geq n$. If A is non–singular (i.e. all $\sigma_i \neq 0$), then using "Pidgin" notation
$$x = \text{“}A^{-1}b\text{“} = \text{“}\big(U\Sigma V^T\big)^{-1}b\text{“} = V\Sigma^{-1}U^Tb$$
with $\Sigma^{-1} = \operatorname{diag}\big(\sigma_1^{-1}, \ldots, \sigma_n^{-1}\big)$. Please note that the terms in quotation marks are not actually defined — the inverse of a non-quadratic matrix **does not exist**! This only serves as a motivation for the actual, valid formula
$$x = V\Sigma^{-1}U^Tb.$$

Using the columns u_i of U and v_i of V, this can be written as

$$x = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i.$$

If A is singular, we can just leave out the terms with $\sigma_i = 0$, yielding

$$x^* = \sum_{i=1}^{\text{rank}(A)} \frac{u_i^T b}{\sigma_i} v_i.$$

This is in fact the least squares solution $x^* = \min_x \|Ax - b\|$ with $\|x^*\|$ minimal (see next section). This also ties in with the pseudo inverse (see also next section), as we have

$$A^+ = \sum_{i=1}^{\text{rank}(A)} v_i \sigma_i^{-1} u_i^T.$$

If A is non–singular, but ill–conditioned (as often the case in tomography), we can calculate the **truncated SVD** solution to stabilize the solution approximation $\hat{x}$,

$$\hat{x} = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i$$

with arbitrary $k \in \{1, \ldots, n\}$.

- In a tomographic setup with $Ax = m$ (see section 1.7) and the SVD of $A = U\Sigma V^T$, we can write

  $$Ax = m \quad \Longleftrightarrow \quad U\Sigma V^T x = m \quad \Longleftrightarrow \quad \Sigma V^T x = U^T m.$$

  We can interpret this as: the columns of U are the detection space modes of A, the columns of V are the image space modes of A. The singular values of A then specify the degree to which a given image space mode is coupled to the corresponding detection space mode. In other words: the singular values of A describe how effectively a given image space mode can be detected by the setup.

- Principle Component Analysis (PCA) $\longrightarrow$ exercise.

- Problem in practice: computing the full SVD is very costly (on the order of $4m^2n+22n^3$) for big matrices and requires a lot of memory.

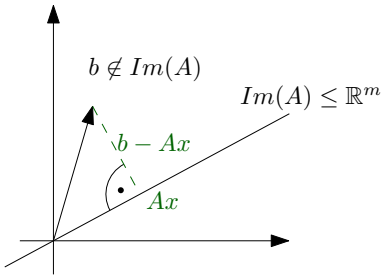
1.11 Least squares problems

Definition (Least squares solution)**.** *Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ with $m \geq n$. A solution $x \in \mathbb{R}^n$ to the minimization problem*

$$\min_x \|Ax - b\| \qquad \text{or} \qquad \min_x \|Ax - b\|^2$$

is called ***least squares solution****.*

Scenario: Suppose we have a linear system we need to solve, for example a linear regression problem (see exercises) or a tomographic problem using the series expansion approach (see previous section). In many cases we have an overdetermined linear system, potentially with noise/errors, which basically always means that there is no solution, as $b \notin \mathrm{Im}(A)$.



Instead of solving $Ax = b$, we now try instead to minimize the error $\|b - Ax\|$. To have $\|b - Ax\|$ minimal, we obviously need $b - Ax \perp \mathrm{Im}(A)$.

Theorem. *Let U be a subspace of $\mathbb{R}^n$, let*

$$U^\perp := \left\{x \in \mathbb{R}^n : \ \langle x, u\rangle = 0 \ \forall u \in U\right\}$$

be the orthogonal complement of U. Let $u \in U$, then we have for all $x \in \mathbb{R}^n$

$$\|x - u\| = \min_{u' \in U} \|x - u'\| \qquad \Longleftrightarrow \qquad x - u \in U^\perp.$$

Proof: Assume $u \in U$ such that $x - u \in U^\perp$, then $\forall u' \in U$:

$$\begin{aligned} \|x - u'\|^2 &= \left\|(x-u) + (u-u')\right\|^2 = \left\langle (x-u) + (u-u'), (x-u) + (u-u')\right\rangle \\ &= \langle x-u, x-u\rangle + 2\underbrace{\langle x-u, u-u'\rangle}_{=0} + \langle u-u', u-u'\rangle \\ &= \|x-u\|^2 + \|u-u'\|^2 \ \geq \ \|x-u\|^2. \end{aligned}$$

To have $\|x - u'\|^2$ minimal we thus require $\|u - u'\|^2 = 0 \ \Leftrightarrow \ u = u'$. □

The unique solution to $\min_{u' \in U} \|x - u'\|$ is called the **orthogonal projection** of x on U. To compute a least squares solution, we thus project b onto $\mathrm{Im}(A)$ orthogonally and solve the linear system there.

Theorem (Normal equation)**.** *Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \geq n$. $x \in \mathbb{R}^n$ is a solution to $\min_x \|Ax - b\|$ if and only if*

$$A^T Ax = A^T b \qquad \textit{(**normal equation**)}.$$

In particular, the normal equation is uniquely solvable if rank$(A) = n$.

Proof: Using the previous theorem we have

$$\begin{aligned} \|b - Ax\| = \min \ &\Longleftrightarrow \ \langle b - Ax, Ax'\rangle = 0 \qquad \forall x' \in \mathbb{R}^n \\ &\Longleftrightarrow \ \left\langle A^T(b - Ax), x'\right\rangle = 0 \qquad \forall x' \in \mathbb{R}^n \\ &\Longleftrightarrow \ A^T(b - Ax) = 0 \qquad \Longleftrightarrow \qquad A^T Ax = A^T b. \end{aligned}$$

A^TA invertible $\Leftrightarrow$ $\operatorname{rank}(A^TA) = n$ $\Leftrightarrow$ $\operatorname{rank}(A) = n$. $\square$

Thus if the normal equation is fulfilled, i.e. $A^TAx = A^Tb$, then we can compute a least squares solution $x = (A^TA)^{-1}A^Tb$. The matrix $(A^TA)^{-1}A^T$ is used quite often, thus it has a name:

Definition (Moore–Penrose inverse)**.** *For $A \in \mathbb{R}^{m\times n}$, $m \geq n$, with linearly independent columns, the unique matrix*

$$A^+ := (A^TA)^{-1}A^T$$

*is called the **Moore–Penrose inverse** of A, or **pseudo inverse** of A.*

If A is invertible (which implies square), then we have $A^+ = A^{-1}$.

Finally, there is another method to solve least squares problems, this time using one of the matrix decompositions:

Theorem (Golub's method)**.** *Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, with rank$(A) = n$ and $b \in \mathbb{R}^m$. If $Q \in \mathbb{R}^{m\times m}$ orthogonal with*

$$Q^TA = \begin{pmatrix} R \\ 0 \end{pmatrix} \qquad \text{and} \qquad Q^Tb = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

with $b_1 \in \mathbb{R}^n$, $b_2 \in \mathbb{R}^{m-n}$ and $R \in \mathbb{R}^{n\times n}$ upper triangular (i.e. from the QR decomposition of A), then

$$x = R^{-1}b_1$$

is a solution to the least squares problem $\min_x \|Ax - b\|$.

In practice we solve $Rx = b_1$ via backward substitution.

Proof: For all $x \in \mathbb{R}^n$:

$$\begin{aligned} \|b - Ax\|^2 &= \left\|Q^T(b - Ax)\right\|^2 = \left\|Q^Tb - Q^TAx\right\|^2 = \left\| \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} - \begin{pmatrix} R \\ 0 \end{pmatrix} x \right\|^2 \\ &= \|b_1 - Rx\|^2 + \|b_2\|^2 \;\geq\; \|b_2\|^2. \end{aligned}$$

As $\operatorname{rank}(A) = n = \operatorname{rank}(R)$ we have R invertible and $b_1 - Rx = 0$ for $x := R^{-1}b_1$, hence $\|b - Ax\| = \min$. $\square$

Remark: Unfortunately $\|b - Ax\| = \min$ generally does not imply $b - Ax = 0$. Hence we define

$$r := b - Ax,$$

the so-called **residual**. Here: $\|r\| = \|b_2\|$.

1.12 Eigenvalue problems

Definition (Eigenvalue, Eigenvector)**.** *Let $A \in \mathbb{R}^{n\times n}$. A vector $x \in \mathbb{C}^n$, $x \neq 0$, is called **eigenvector** of A if*

$$Ax = \lambda x$$

*for some $\lambda \in \mathbb{C}$. The $\lambda \in \mathbb{C}$ is then called **eigenvalue** of A.*

Theorem (Eigenvalues of symmetric matrices)**.** *Let $A \in \mathbb{R}^{n\times n}$, A symmetric ($A = A^T$). Then all eigenvalues λ of A are real, $\lambda \in \mathbb{R}$.*

How to calculate eigenvalues? $Ax = \lambda x \;\Rightarrow\; (\lambda I_n - A)x = 0$, a linear system. We know $Bx = 0$ has only the solution $x = 0 \;\Leftrightarrow\; B$ invertible $\;\Leftrightarrow\; \det(B) \neq 0$. We want a solution $x \neq 0$, thus $\det(\lambda I_n - A) = 0$ — this is called the **characteristic equation** of A. We can write this as

$$\det(\lambda I_n - A) = \lambda^n + c_1\lambda^{n-1} + \ldots + c_n\lambda^0,$$

the **characteristic polynomial** of A. The zero–crossings of the characteristic polynomial of A are the eigenvalues λ_i of A (hence we can have at most n distinct eigenvalues due to the fundamental theorem of algebra), the eigenvectors x_i for each λ_i can then be computed by solving $(\lambda_i I_n - A)x_i = 0$ for x_i.

Example: $A = \begin{pmatrix} -1 & \sqrt{2} \\ -\sqrt{2} & 1 \end{pmatrix}$.

$$\begin{aligned} \det(\lambda I_2 - A) &= \det \begin{pmatrix} \lambda+1 & -\sqrt{2} \\ \sqrt{2} & \lambda-1 \end{pmatrix} = (\lambda+1)(\lambda-1) + 2 = \lambda^2 - 1 + 2 = \lambda^2 + 1 \\ &= (\lambda + i)(\lambda - i), \end{aligned}$$

thus $\lambda_1 = i$, $\lambda_2 = -i$.

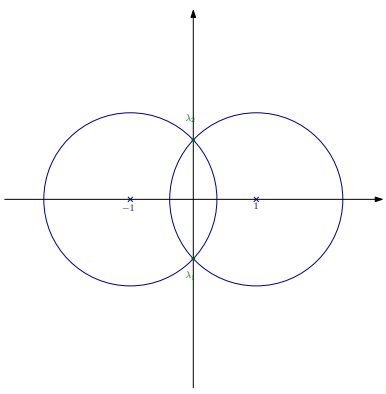
Properties: Let $A \in \mathbb{R}^{n\times n}$.

- If A diagonal $\left(\begin{smallmatrix} \diagdown \end{smallmatrix}\right)$ or A triangular $\left(\begin{smallmatrix} \diagdown \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} \diagdown \end{smallmatrix}\right)$, then the eigenvalues are the diagonal entries of A.
- A invertible $\;\Longleftrightarrow\;$ $\lambda = 0$ is **no** eigenvalue of A.

Theorem (Gershgorin)**.** *Let $A \in \mathbb{R}^{n\times n}$. Any eigenvalue λ of A is located in one of the closed discs on the complex plane centered at a_{ii} with radius $\sum_{j=1,i\neq j}^{n} |a_{ij}|$. In short: for any eigenvalue λ of A we have*

$$\exists i \in \{1, \ldots, n\} : \qquad |\lambda - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

Gershgorin discs from previous example:



Definition (Multiplicity)**.** *Let $A \in \mathbb{R}^{n\times n}$ and let $\lambda_0 \in \mathbb{C}$ be an eigenvalue of A.*

- *The number of factors $(\lambda - \lambda_0)$ in the characteristic polynomial of A is called* ***algebraic multiplicity*** *of λ_0.*
- *The eigenvectors of λ_0 span a subspace of $\mathbb{C}^n$ (the* ***eigenspace*** *of λ_0). The dimension of the eigenspace of λ_0 is called* ***geometric multiplicity*** *of λ_0.*

Examples:

- previous A: λ_1, λ_2 each have algebraic multiplicity 1.
- $B = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ is diagonal and has eigenvalues $\lambda_1 = \lambda_2 = 1$ with multiplicity 2. For the eigenvectors we have $Bx = x \ \forall x \in \mathbb{R}^2$ and thus eigenvectors $\left(\begin{smallmatrix} s \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 \\ t \end{smallmatrix}\right)$ with $s, t \in \mathbb{R}$, that is geometric multiplicity 2.
- $C = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ is triangular and has eigenvalues $\lambda_1 = \lambda_2 = 1$ with multiplicity 2. For the eigenvectors we solve $(\lambda_1 I_2 - C)x = \left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) x \stackrel{!}{=} 0$, yielding $x = \left(\begin{smallmatrix} t \\ 0 \end{smallmatrix}\right)$ for $t \in \mathbb{R}$, that is geometric multiplicity 1.

Definition (Similarity)**.** *Let $A, B \in \mathbb{R}^{n\times n}$. A and B are called* ***similar****, if there exists $P \in \mathbb{R}^{n\times n}$ invertible such that*

$$B = P^{-1}AP.$$

Properties: Let $A, B \in \mathbb{R}^{n\times n}$.

- Let $P \in \mathbb{R}^{n\times n}$ invertible. The transformation $P^{-1}AP$ is a change of basis for the linear mapping f with $f(x) = Ax$.
- If A, B are similar, then they have the same characteristic polynomial, eigenvalues and algebraic and geometric multiplicity. Rank and determinant are also the same.

Definition (Diagonalization)**.** *Let $A \in \mathbb{R}^{n\times n}$. If there exists $P \in \mathbb{R}^{n\times n}$ invertible such that $P^{-1}AP = \begin{pmatrix} \ddots \end{pmatrix}$, then A is called* ***diagonalizable****.*

Properties: Let $A \in \mathbb{R}^{n\times n}$.

- Let $\lambda \in \mathbb{C}$ be an eigenvalue of A. Then geometric multiplicity of $\lambda \leq$ algebraic multiplicity of λ. In case of $\lneq$, λ as well as A are called **defective**.
- A diagonalizable $\iff$ A has n linearly independent eigenvectors (they form a basis)
  $\iff$ geometric = algebraic multiplicity for all eigenvalues of A
  (or in short: A non–defective).
- A symmetric ($A = A^T$). Then all eigenvalues of A are real, A is diagonalizable and the P from $P^{-1}AP = \begin{pmatrix} \ddots \end{pmatrix}$, is orthogonal.
- A normal ($\Leftrightarrow$ $A^TA = AA^T$) $\iff$ A diagonalizable with P orthogonal.

Practial approaches to calculating eigenvalues / eigenvectors. For the remainder of this section we assume $A \in \mathbb{R}^{n\times n}$ symmetric ($A = A^T$) for simplicity.

If we know an eigenvector $x \in \mathbb{R}^n$ of A (since A is symmetric, $\mathbb{R}^n$ is enough now), to find the corresponding eigenvalue $\lambda \in \mathbb{R}$, we can study the least squares problem $\|Ax - \lambda x\| = \min$ by interpreting it as

$$\|x\lambda - Ax\| = \min_{\lambda},$$

where λ is the unknown, x the $n \times 1$ matrix and Ax the right hand side. Using the normal equation we have

$$x^T x\lambda = x^T Ax$$

and thus

$$\lambda = \frac{x^T Ax}{x^T x} =: r(x).$$

$r(x)$ is also called the **Rayleigh quotient**.

To calculate an eigenpair of A we can use the

Rayleigh quotient iteration

Input: x^0 estimate of an eigenvector of A
Output: (x, λ) eigenpair of A

$x^0 = \frac{x^0}{\|x^0\|}$
$\lambda^0 = (x^0)^T Ax^0$ (Rayleigh quotient)
for $k = 1, 2, \ldots$
solve $(A - \lambda^{k-1} I_n)y = x^{k-1}$ for y
$x^k = \frac{y}{\|y\|}$
$\lambda^k = (x^k)^T Ax^k$ (Rayleigh quotient)

This algorithm converges cubically for almost all inputs. In practice A is reduced to tridiagonal form beforehand to speed up computations. **Deflation**, i.e. $A' := A - \lambda xx^T$ allows to calculate other / all eigenpairs.

A better method for this is the

QR Algorithm

Input: A
Output: diagonalized A with (orthonormal) eigenvectors in Q

$A^0 = A$
for $k = 1, 2, \ldots$
$Q^k R^k = A^{k-1}$ (QR factorization of A^{k-1})
$A^k = R^k Q^k$

The trick here is that each A^k is similar to A^{k-1}, as

$$A^k = R^k Q^k = (Q^k)^T Q^k R^k Q^k = (Q^k)^T A^{k-1} Q^k,$$

which iteratively yields $A^k = Q^T A^0 Q$ with $Q := Q^1 Q^2 \cdots Q^k$.

In practice A is reduced to tridiagonal form beforehand, a shifting strategy and deflation are added to achieve fast convergence.

There are of course multiple other methods available to calculate eigenvalues or eigenpairs (for example the power method or the divide–and–conquer method).

Application examples:

- Principal Component Analysis (PCA) $\longrightarrow$ exercise.
- Google Page Rank.
- Rotation matrices. An axis of rotation v is not changed by a rotation R, i.e. $Rv = v \Leftrightarrow (R - I)v = 0$. Thus: the eigenvector of R corresponding to the eigenvalue 1 is the axis of rotation.

  Example: Rotation in $\mathbb{R}^3$ around x–axis with angle α:

$$R_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix}$$

2 Analysis

2.1 Metric spaces

Arbitrary sets M can be equipped with a notion of "distance" via a metric.

Definition (Metric)**.** *Let M be a set, then a mapping $d : M \times M \to \mathbb{R}_0^+$ is called* ***metric on*** *M, if $\forall x, y, z \in M$ we have*

(1) $d(x, y) = 0 \iff x = y$

(2) $d(x, y) = d(y, x)$

(3) $d(x, z) \leq d(x, y) + d(y, z)$.

The pair (M, d) is then called a ***metric space****.*

Examples:

- Let $G = (V, E)$ be an undirected, connected graph with weights $w : E \to \mathbb{R}^+$, then for two vertices $x, y \in V$

  $$d(x, y) := \text{ "length of shortest path connecting } x \text{ and } y\text{"}$$

  is a metric.

- Let $(V, \|\cdot\|)$ be a normed linear space. Then

  $$d(x, y) := \|y - x\|$$

  for $x, y \in V$ defines a metric.

  For the special case $V = \mathbb{R}^n$, there are two notable special cases:

  - p**–norm:** For $1 \leq p < \infty$ the p–norm of $x \in \mathbb{R}^n$ is defined as

    $$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}},$$

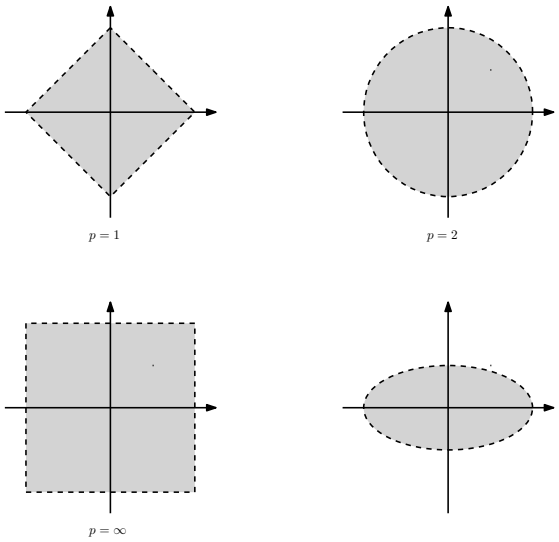
    and for $p = \infty$ as

    $$\|x\|_\infty := \max_{i=1,\ldots,n} |x_i|.$$

  - **energy norm:** For $A \in \mathbb{R}^{n \times n}$ positive definite, the energy norm of $x \in \mathbb{R}^n$ is defined as

    $$\|x\|_A := \sqrt{x^T A x}.$$

Looking at the open unit ball $B_1(0) := \left\{ x \in \mathbb{R}^n : \|x\| < 1 \right\}$ helps to get a feel for these norms. For example in the case of $n = 2$:



Open unit ball for the various norms.
The lower right entry is the energy norm for $A = \left(\begin{smallmatrix} 0.5 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

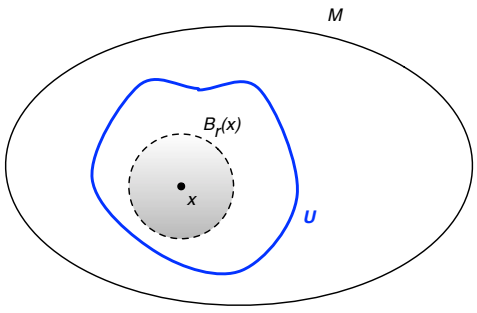
2.2 Topology in metric spaces

Definition (Open ball, neighborhood)**.** *Let* (M, d) *be a metric space. For* $x \in M$ *and* $r > 0$

$$B_r(x) := \{y \in M : \ d(x, y) < r\}$$

is called the ***open ball*** *around* x *with radius* r*.*
A set $U \subset M$ *is called* ***neighborhood*** *of* $x \in M$*, if there exists* $\varepsilon > 0$ *such that*

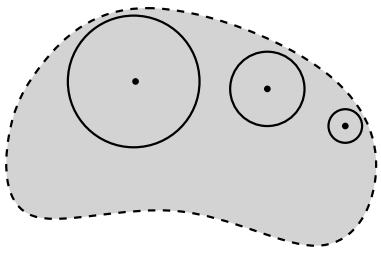
$$x \in B_\varepsilon(x) \subset U.$$



Definition (Open, closed sets)**.** *Let* (M, d) *be a metric space. A subset* $U \subset M$ *is called* ***open****, if* U *is a neighborhood of every* $x \in U$*, that is:*

$$\forall x \in U \quad \exists \varepsilon > 0 : \qquad x \in B_\varepsilon(x) \subset U.$$

A subset $A \subset M$ *is called* ***closed****, if the complement* $A^c = M \setminus A$ *is open.*



Remarks: Let (M, d) be a metric space.

- The set $\mathcal{T}$ of all open subsets of (M, d) is called **topology** on M.
- $\emptyset$ and M are both open **and** closed!
- If the subsets U, V of M are open / closed, then $U \cup V$, $U \cap V$ are open / closed.
- If a set $U \subset M$ is open or closed depends both on M and d. For example $[0, 1]$ is not open in $(\mathbb{R}, |\cdot|)$, but it is open in $([0, 1], |\cdot|)$. It will turn out however, that all norms of $\mathbb{R}^n$ induce the same topology.

Examples: Let (M, d) be a metric space.

- $\{x\}$ for $x \in M$ is closed.
- Let $x \in M$, $r > 0$.
  - $B_r(x) := \{y \in M : \ d(x, y) < r\}$ is open (**open ball** around x with radius r).
  - $K_r(x) := \{y \in M : \ d(x, y) \leq r\}$ is closed (**closed ball** around x with radius r).

  Special case $(\mathbb{R}, |\cdot|)$: intervals $(a, b) \subset \mathbb{R}$ are open, intervals $[a, b] \subset \mathbb{R}$ are closed.

Definition (Boundary)**.** *Let (M, d) be a metric space, $A \subset M$ a subset. $x \in M$ is called a **boundary point** of A, if for all $\varepsilon > 0$ $B_\varepsilon(x)$ contains both a point of A and $M \setminus A$. The set*

$$\partial A := \{x \in M : \ x \text{ boundary point of } A\}$$

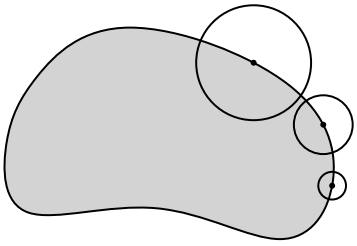
*is called **boundary** of A. The set*

$$int(A) := A \setminus \partial A$$

*is called **interior** of A, and the set*

$$\overline{A} := A \cup \partial A$$

*is called the **closure** of A.*



Remarks: Let (M, d) be a metric space, $A \subset M$.

- $\operatorname{int}(A)$ is open, $\overline{A}$ is closed, ∂A is closed.
- $\operatorname{int}(A) \subset A \subset \overline{A}$.
- $\partial A = \overline{A} \setminus \operatorname{int}(A)$.
- A closed $\iff \quad A = \overline{A}$.

Examples: Let (M, d) be a metric space.

- Let $x \in M$, $r > 0$.
  - $S_r(x) := \{y \in M : \ d(x, y) = r\}$ is closed (**sphere** around x with radius r).
  - $S_r(x) = \partial K_r(x) = \partial B_r(x) = K_r(x) \setminus B_r(x)$.
- Special case $(\mathbb{R}, |\cdot|)$:
  Let $(a, b) \subset A \subset [a, b] \subset \mathbb{R}$, then $\partial A = \{a, b\}$, $\operatorname{int}(A) = (a, b)$ and $\overline{A} = [a, b]$.
  Also $\partial\mathbb{Q} = \overline{\mathbb{Q}} = \mathbb{R}$, $\operatorname{int}(\mathbb{Q}) = \emptyset$, but $\overline{\mathbb{Q}} \neq \overline{\operatorname{int}(\mathbb{Q})}$.

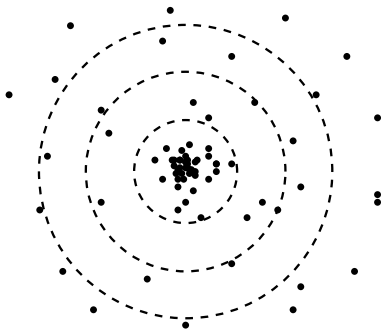
2.3 Convergence and compactness

Definition (Convergence)**.** *Let (M, d) be a metric space. A sequence $(x_k)_{k\in\mathbb{N}} \subset M$ is called **convergent** to a **limit** $a \in M$, if for every neighborhood U of a there exists a $N \in \mathbb{N}$ such that $x_k \in U \ \forall k \geq N$. We write in short*

$$\lim_{k\to\infty} x_k = a \qquad \textit{or} \qquad x_k \overset{k\to\infty}{\longrightarrow} a.$$

Equivalently we have:

$$\begin{aligned} x_k \overset{k\to\infty}{\longrightarrow} a \quad &\iff \quad \lim_{k\to\infty} d(x_k, a) = 0 \\ &\iff \quad \forall \varepsilon > 0 \ \exists N \in \mathbb{N} \textit{ such that } d(x_k, a) < \varepsilon \ \forall k \geq N. \end{aligned}$$



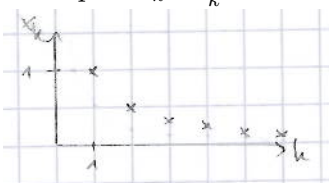
Remarks: Let (M, d) be a metric space.

- If $(x_k)_{k\in\mathbb{N}}$ is a convergent sequence in M, the limit $\lim_{k\to\infty} x_k$ is uniquely defined, and every subsequence of (x_k) converges to it.

- Let $A \subset M$. Then:

$$A \text{ closed} \quad \Longleftrightarrow \quad \text{every convergent sequence } (x_k) \subset A \text{ has } \lim_{k\to\infty} x_k \in A$$

- Example: $x_k = \frac{1}{k}$ for $k \in \mathbb{N}$. Then $\lim_{k\to\infty} x_k = 0$.



Definition. *Let (M, d) be a metric space. A sequence $(x_k)_{k\in\mathbb{N}} \subset M$ is called **Cauchy sequence**, if $\forall \varepsilon > 0\ \exists N \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon\ \forall n, m \geq N$.*

*(M, d) is called **complete**, if every Cauchy sequence converges.*

Remark: every convergent sequence is a Cauchy sequence.

Examples:

- $(\mathbb{R}, |\cdot|)$ is complete.
- $(\mathbb{Q}, |\cdot|)$ is not complete, as for example $x_1 := 1$, $x_{k+1} := \frac{x_k}{2} + \frac{1}{x_k}$ has the limit $\lim_{k\to\infty} x_k = \sqrt{2} \notin \mathbb{Q}$.
- A normed linear space $(V, \|\cdot\|)$ which is complete is also called **Banach space**.
- A linear space V with a scalar product $\langle\cdot,\cdot\rangle$ induces a metric via the induced norm $\|x\| := \sqrt{\langle x, x\rangle}$ for $x \in V$. If V is complete it is also called **Hilbert space**.

$$\begin{array}{ccc} \text{linear space with } \langle\cdot,\cdot\rangle & \overset{\text{complete}}{\longrightarrow} & \text{Hilbert space} \\ \downarrow & & \downarrow \\ \text{linear space with } \|\cdot\| & \overset{\text{complete}}{\longrightarrow} & \text{Banach space} \\ \downarrow & & \downarrow \\ \text{metric space with } d(\cdot,\cdot) & \overset{\text{complete}}{\longrightarrow} & \text{complete metric space} \end{array}$$

Banach/Hilbert space examples:

- $(\mathbb{R}^n, \|\cdot\|_p)$ is a Banach space. For $p = 2$ we have $\|x\|_2 = \sqrt{\langle x, x\rangle}$, hence $(\mathbb{R}^n, \|\cdot\|_2)$ is a Hilbert space.
- For X set and $p \geq 1$

$$\mathcal{L}^p(X) := \left\{ f : X \to \mathbb{R} : \int |f(x)|^p \, dx < \infty \right\}$$

  is a linear space using Lebesgue integration.

$$\|f\|_p := \left(\int |f(x)|^p \, dx \right)^{\frac{1}{p}}$$

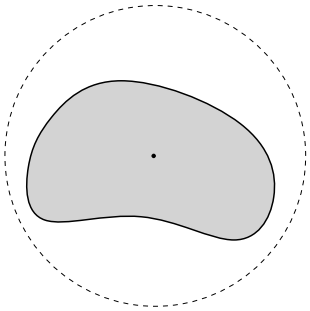
however is not a norm, just a semi–norm, as there exist functions $f \neq 0$ with $\|f\|_p = 0$ (for example with $X = \mathbb{R}$: $f(x) = 0$ for $x \neq 0$, and $f(0) = 1$). Using

$$f \text{ equivalent to } g \qquad \Longleftrightarrow \qquad \|f - g\|_p = 0,$$

we can define $L^p(X)$ as the equivalence classes of this equivalence relation. $\big(L^p(X), \|\cdot\|_p\big)$ is then a Banach space, for $p = 2$ it is also a Hilbert space.

Definition (Bounded set)**.** *Let (M, d) be a metric space. A subset $U \subset M$ is called* ***bounded*** *if $\sup_{x,y\in U} d(x, y) < \infty$. Alternatively:*

$$U \text{ is bounded} \qquad \Longleftrightarrow \qquad \exists \varepsilon > 0, x \in U \quad \text{such that} \quad U \subset B_\varepsilon(x).$$



Definition (Compactness)**.** *Let (M, d) be a metric space. (M, d) is called* ***compact****, if every sequence $(x_k)_{k\in\mathbb{N}} \subset M$ has a convergent subsequence.*
A subset $U \subset M$ is called ***compact****, if every sequence $(x_k)_{k\in\mathbb{N}} \subset U$ has a convergent subsequence with a limit in U.*

Theorem (Bolzano–Weierstraß)**.** *Consider $(\mathbb{R}^n, \|\cdot\|_\infty)$ and $U \subset \mathbb{R}^n$. Then*

$$U \text{ compact} \quad \Longleftrightarrow \quad U \text{ bounded and closed in } \mathbb{R}^n.$$

Examples:

- $[a, b] \subset \mathbb{R}$ is compact in $(\mathbb{R}, |\cdot|)$.
- $K_r(x)$ is compact in $\mathbb{R}^n$ (for $x \in \mathbb{R}^n$, $r > 0$).

Theorem (Equivalence of norms in $\mathbb{R}^n$)**.** *For every norm $\|\cdot\|$ on $\mathbb{R}^n$ there exist $c_1, c_2 > 0$ with*

$$c_1\|x\|_\infty \leq \|x\| \leq c_2\|x\|_\infty \qquad \forall x \in \mathbb{R}^n.$$

In other words: For $\mathbb{R}^n$, properties like convergence, Cauchy sequence, completeness, open, closed, bounded, compact, boundary, interior, closure do not depend on the norm inducing the metric! This is in fact true for any finite dimensional normed linear space!

Remark (Peculiarities of infinite dimensional spaces). The reason why one has to emphasize the fact whether a space is finite dimensional or not, is that a lot of concepts and theorems which hold true in finite dimensional spaces cannot be carried over to infinite dimensional

spaces. A famous example is the theorem of Bolzano-Weierstraß:
Let us consider the space of all bounded sequences

$$\ell^\infty = \left\{x = (x_n)_{n\in\mathbb{N}} : x_n \in \mathbb{R}, \|x\|_\infty < \infty\right\},$$

where

$$\|x\|_\infty = \sup_{n\in\mathbb{N}} |x_n| .$$

The closed unit ball in ℓ^∞, i.e.,

$$K_1(0) = \left\{x = (x_n)_{n\in\mathbb{N}} : \|x\|_\infty \leq 1\right\} \subset \ell^\infty$$

is obviously bounded and closed. However, considering the sequence (of sequences) $(x^k)_{k\in\mathbb{N}} \subset K_1(0)$ which is given by:

$x^1 = (1, 0, 0, 0, \ldots)$, $x^2 = (0, 1, 0, 0, \ldots)$, $x^3 = (0, 0, 1, 0, \ldots)$, $x^4 = (0, 0, 0, 1, \ldots)$, etc.

one realizes that there exists no converging subsequence of this sequence. Hence, $K_1(0)$ cannot be compact!
It is actually even worse. It can be shown that it is possible to construct such sequences, i.e. sequences with no convergent subsequence, in *any* infinite dimensional normed space!

2.4 Continuity

Remark (Motivation). The fact whether a function is continuous or not has several theoretical and practical implications. Probably the most popular one is that a continuous function is compatible with limits which means (roughly speaking) that for any converging sequence $(x_n)_{n\in\mathbb{N}}$

$$f(\lim_{n\to\infty} x_n) = \lim_{n\to\infty} f(x_n)$$

holds true. A sophisticated example where continuity is important are *variational approaches* for image processing and computer vision (e.g., CAMP II lecture on variational image segmentation) for finding minimizers of a functional, i.e., a function mapping other functions to real values - a typical scenario looks like follows:
Let $\Omega \subset \mathbb{R}^n$ be a bounded set and $\mathcal{F}(\Omega)$ some space of functions defined on Ω. Let further $E : \mathcal{F}(\Omega) \to \mathbb{R}$ denote the functional - often called energy - to be minimized:

$$\min_{u\in\mathcal{F}(\Omega)} E(u).$$

Besides the computation of the so-called *first variation* (see CAMP II for instance) one is often interested in certain properties of E, such as its *continuity*:
Let $(u_n)_{n\in\mathbb{N}} \subset \mathcal{F}(\Omega)$ be a sequence of functions converging w.r.t. the topology induced by the norm of $\mathcal{F}$, e.g., let $\mathcal{F}(\Omega) = L^2(\Omega)$ and

$$u_n \to u^* \in L^2(\Omega).$$

If E is continuous, i.e.,

$$\lim_{n\to\infty} E(u_n) = E(u^*)$$

holds for any sequence in L^2 converging to u^*, we may also draw the following conclusion: If u^* is a solution of the minimization problem above and E is continuous in u^* (and hopefully in a small vicinity of it, too) we may conclude that solving the minimization problem above is also numerically feasible. The reason is that, given a suitable numerical method for minimizing E, an initial numerical guess $\hat{u}_0$, which is sufficiently close to u^*, may lead to a sequence of iterates $\hat{u}_1$, $\hat{u}_2$, $\hat{u}_3$, etc., such that

$$\hat{u}_n \to \hat{u}^* \quad \text{, and } \|\hat{u}^* - u^*\|_{L^2} \leq \delta,$$

where δ is some sufficiently small tolerance we are willing to accept. However, the fact of E being continuous is often not very obvious, because E is typically of the form

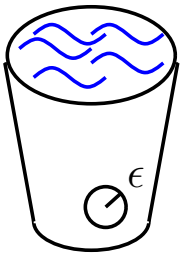
$$E(u) = \int_\Omega F(u)\, dx.$$

Even in the case $F = id$ one would have to prove that

$$\lim_{n\to\infty} \int_\Omega u_n(x)\, dx = \int_\Omega u^*(x)\, dx$$

for any sequence converging to u^*. Noting that the integral bears a hidden limit process, too, one ends up with the question if switching these two limit processes is allowed. In this particular case the answer will be given a bit later, but switching the order of limits is not always possible as the following example shows.

Example (The Leaky Water Bucket):



Consider a filled water bucket with maximum volume $5l$ and a hole of radius ϵ. Let further $V(\epsilon, t)$ denote the volume of water in the bucket at time t. Wondering what happens, if one waits forever, but letting the hole shrink at the same time, one has the following two possibilities:

- First closing the hole and waiting forever, i.e.,

$$\lim_{t\to\infty} \lim_{\epsilon\to 0} V(\epsilon, t) = 5l,$$

- or waiting forever and closing the hole afterwards, i.e.,

$$\lim_{\epsilon\to 0} \lim_{t\to\infty} V(\epsilon, t) = 0l.$$

This example shows that exchanging the order of limits is not always possible!

Definition (Continuity)**.** *Let (M, d_1), (N, d_2) be metric spaces, let $U \subset M$ and $f : U \to N$. If for every sequence $(x_k)_{k\in\mathbb{N}} \subset U$ with $\lim_{k\to\infty} x_k = a \in \overline{U}$ we have $\lim_{k\to\infty} f(x_k) = c \in N$, then c is called **limit** of f in a. In short: $\lim_{x\to a} f(x) = c$ or $f(x) \xrightarrow{x\to a} c$.*

*f is called **continuous in** a, if $\lim_{x\to a} f(x) = f(a)$.*

*f is called **continuous** on $V \subset U$, if f is continuous for every $a \in V$.*

Properties:

- Most "regular" combinations of continuous functions f, g are continuous again, that is $f \circ g$, fg, $\frac{f}{g}$ (for $g \neq 0$), $f + g$, $f - g$ are continuous if f, g (operating on appropriate sets) are continuous.
- Let (M, d) be a metric space, let $f : M \to \mathbb{R}^n$ with $f(x) = \big(f_1(x), \ldots, f_n(x)\big)^T$. Then

$$f \text{ is continuous in } a \in M \quad \Longleftrightarrow \quad f_i : M \to \mathbb{R} \text{ is continuous in } a \quad \forall i = 1, \ldots, n.$$

Examples:

- $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = c$ for $x \in \mathbb{R}^n$, $c \in \mathbb{R}$, is continuous.
- $p_i : \mathbb{R}^n \to \mathbb{R}$, $p_i(x_1, \ldots, x_n) = x_i$ for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $i = 1, \ldots, n$, is continuous.
- Let I be a finite subset of $\prod_{i=1}^n \mathbb{N}_0$, then $f : \mathbb{R}^n \to \mathbb{R}$ with

$$f(x_1, \ldots, x_n) = \sum_{\alpha\in I} c_\alpha x_1^{\alpha_1} \cdots x_n^{\alpha_n}$$

  for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $c_\alpha \in \mathbb{R}$, and $\alpha = (\alpha_1, \ldots, \alpha_n) \in I$, is called **polynomial** in n variables of degree $N = \max\{\sum_{i=1}^n \alpha_i : \ \alpha \in I\}$; all polynomials are continuous.
- Every linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous, as using the associated matrix $A = (a_{ji}) \in \mathbb{R}^{m\times n}$ it can be written as $f(x) = \big(f_1(x), \ldots, f_m(x)\big)^T$ with

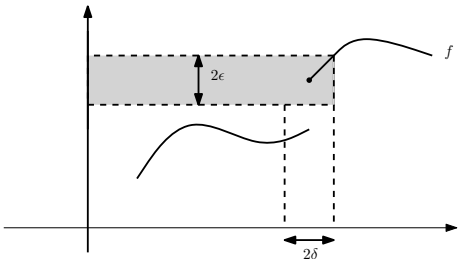
$$f_j(x) = \sum_{i=1}^n a_{ji} x_i.$$

Often, several (equivalent) characterizations of a mathematical property are available. This is also true for continuity. Having several characterizations at hand provides the flexibility of picking out the most appropriate one, e.g., the easiest to prove, for a particular application.

Theorem (ε–δ–version of continuity)**.** *Let (M, d_1), (N, d_2) be metric spaces, $f : M \to N$. Then*

f continuous in $a \in M$ $\quad\Longleftrightarrow\quad$ $\forall \varepsilon > 0\ \exists \delta > 0$ such that $d_1(x, a) < \delta \ \Rightarrow\ d_2\big(f(x), f(a)\big) < \varepsilon$.

Example of a function f that is not continuous:



Theorem (Version for normed linear spaces / linear mappings)**.** *Let* $(V, \|\cdot\|_V)$, $(W, \|\cdot\|_W)$ *be normed linear spaces, let* $f : V \to W$ *be a linear mapping. Then:*

$$f \text{ continuous} \quad \Longleftrightarrow \quad \exists C > 0 \text{ such that } \|f(x)\|_W \leq C\|x\|_V \quad \forall x \in V.$$

Example (continued):
Consider the variational problem

$$\min_{u \in L^2(\Omega)} \int_\Omega u \, dx$$

again. At first, one can show that

$$\int_\Omega v(x)w(x) \, dx = \langle v(x)w(x) \rangle_{L^2}$$

is an inner/scalar product on L^2. Since for inner products the Cauchy-Schwarz inequality

$$| \langle v, w \rangle_{L^2} | \leq \|v\|_{L^2} \, \|w\|_{L^2}$$

holds, one may conclude that

$$|E(u)| = \left| \int_\Omega u(x) \cdot 1 \, dx \right| \leq \left(\int_\Omega 1^2 \, dx \right)^{\frac{1}{2}} \left(\int_\Omega u(x)^2 \, dx \right)^{\frac{1}{2}} = C \, \|u\|_{L^2} \,,$$

where

$$C = \left(\int_\Omega 1^2 \, dx \right)^{\frac{1}{2}} < \infty,$$

because Ω was chosen to be a bounded domain, and this eventually means that E is continuous!

Theorem (Connection to topology)**.** *Let* (M, d_1), (N, d_2) *be metric spaces,* $f : M \to N$. *Then:*

$$\begin{aligned} f \text{ continuous on } M \quad &\Longleftrightarrow \quad \forall U \subset N \text{ open we have } f^{-1}(U) \subset M \text{ open} \\ &\Longleftrightarrow \quad \forall A \subset N \text{ closed we have } f^{-1}(A) \subset M \text{ closed.} \end{aligned}$$

(For $U \subset N$ *we have* $f^{-1}(U) := \{x \in M : \ f(x) \in U\}$*.)*

Example: Let (M, d) be a metric space, $f : M \to \mathbb{R}$ continuous, $c \in \mathbb{R}$. Then $\{x \in M : f(x) < c\}$, $\{x \in M : f(x) > c\}$ are open, $\{f(x) \geq c\}$, $\{f(x) \leq c\}$, $\{f(x) = c\}$ are closed. The set $\{f(x) = c\}$ is called **level set**.

Theorem. *Let (M, d_1), (N, d_2) be metric spaces, $f : M \to N$ continuous and M compact. Then $f(M)$ is compact.*

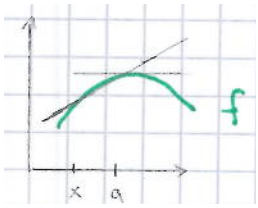
Theorem (Min/Max)**.** *Let (M, d) be a metric space, $f : M \to \mathbb{R}$ continuous and $M \neq \emptyset$ compact. Then $f(M)$ is bounded, and f attains its minimum / maximum on M, that is: $\exists p, q \in M$ such that*

$$f(p) = \max_{x \in M} f(x), \qquad f(q) = \min_{x \in M} f(x).$$

2.5 Differentiability

Given a function $f : \mathbb{R} \to \mathbb{R}$, we want to know the slope of f in $a \in \mathbb{R}$. One way to calculate it is to calculate the difference quotient, that is the slope of the secant line:

$$\Delta f(a; x) = \frac{f(x) - f(a)}{x - a}.$$



Another way to calculate the slope of f is:

Definition (Derivative)**.** *Let $I \subset \mathbb{R}$ be open, $f : I \to \mathbb{R}$. f is called **differentiable** in $a \in I$ if*

$$f'(a) := \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}$$

*exists. $f'(a)$ is called the **derivative** of f in a, also denoted as $\frac{d}{dx} f(a)$.*

Properties:

- Let $I \subset \mathbb{R}$ be open, $f, g : I \to \mathbb{R}$ differentiable in $a \in I$. Then $f + g$, fg and $\frac{f}{g}$ (if $g(a) \neq 0$) are differentiable in a too, and we have

$$\begin{aligned} (f + g)'(a) &= f'(a) + g'(a) \\ (fg)'(a) &= f'(a)g(a) + f(a)g'(a) \qquad \text{(product rule)} \\ \left(\frac{f}{g}\right)'(a) &= \frac{f'(a)g(a) - f(a)g'(a)}{g(a)^2} \qquad \text{(quotient rule)} \end{aligned}$$

- Let $I, J \subset \mathbb{R}$ be open and $f : I \to J$, $g : J \to \mathbb{R}$. If f is differentiable in $a \in I$ and g in $f(a) \in J$, then $g \circ f : I \to \mathbb{R}$ is differentiable in a, and we have

$$(g \circ f)'(a) = g'\big(f(a)\big) \cdot f'(a) \qquad \text{(chain rule)}.$$

Examples:

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = ax^n$ for $a \in \mathbb{R}$, $n \in \mathbb{N}$, then $f'(x) = n \cdot ax^{n-1}$.
- $f : (0, \infty) \to \mathbb{R}$, $f(x) = \ln(x)$, then $f'(x) = \frac{1}{x}$.
- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \exp(x)$, then $f'(x) = \exp(x)$.
- $f, g : \mathbb{R} \to \mathbb{R}$, $f(x) = \sin(x)$, $g(x) = \cos(x)$, then

$$f'(x) = \cos(x), \qquad g'(x) = -\sin(x).$$

Now, on to higher dimensions, that is functions like $f : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^m$!

2.5.1 Partial derivatives

Definition (Directional / partial derivative)**.** *Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}$. If for $x \in \Omega$ and a "direction vector" $v \in \mathbb{R}^n$*

$$\partial_v f(x) := \lim_{\substack{t \to 0, t \neq 0 \\ x+tv \in \Omega}} \frac{f(x+tv) - f(x)}{t}$$

*exists, then $\partial_v f(x)$ is called the **directional derivative** of f in x in direction of v (this is also called the **Gateaux derivative**).*

*Let $\{e_1, \ldots, e_n\}$ be the canonical basis of $\mathbb{R}^n$, then for $v = e_i$, $i \in \{1, \ldots, n\}$, $\partial_v f(x)$ is called the i–**th partial derivative** of f in x, denoted as $\partial_i f(x)$ or $\frac{\partial}{\partial x_i} f(x)$.*

*f is called **partially differentiable** in Ω, if $\partial_i f(x)$ exists for all $x \in \Omega$ and $i = 1, \ldots, n$.*

*Let $f : \Omega \to \mathbb{R}$ be partially differentiable in Ω, then if the $\partial_i f : \Omega \to \mathbb{R}$ are continuous for all $i = 1, \ldots, n$, f is called **continuously differentiable** in Ω.*

Example: $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 \sin(y)$

$$\begin{aligned} \partial_1 f(x,y) &= 2x \sin(y), & \partial_2 \partial_1 f(x,y) &= 2x \cos(y) \\ \partial_2 f(x,y) &= x^2 \cos(y), & \partial_1 \partial_2 f(x,y) &= 2x \cos(y) \end{aligned}$$

Theorem (Schwarz)**.** *Let $\Omega \subset \mathbb{R}^n$ be open, $f : \Omega \to \mathbb{R}$ twice continuously differentiable, then*

$$\partial_i \partial_j f(x) = \partial_j \partial_i f(x)$$

for all $i, j = 1, \ldots, n$ and $x \in \Omega$.

Notation: Let $\Omega \subset \mathbb{R}^n$, $k \in \mathbb{N}$.

$$\begin{aligned} C^0(\Omega) &:= C(\Omega) := \{f : \Omega \to \mathbb{R} : \ f \text{ continuous}\} \\ C^k(\Omega) &:= \{f : \Omega \to \mathbb{R} : \ f \ k\text{–times continuously differentiable}\} \end{aligned}$$

and

$$C^\infty(\Omega) := \bigcap_{k \in \mathbb{N}} C^k(\Omega).$$

We have

$$C^0(\Omega) \supset C^1(\Omega) \supset C^2(\Omega) \supset \cdots \supset C^\infty(\Omega).$$

Definition (Gradient)**.** *Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}$ partially differentiable, then*

$$grad\ f(x) := \nabla f(x) := \big(\partial_1 f(x), \ldots, \partial_n f(x)\big)^T$$

is called ***gradient*** *of f at $x \in \Omega$.*

Properties:

- Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}$ partially differentiable, $v \in \mathbb{R}^n$.

$$\partial_v f(x) = \big\langle \nabla f(x), v \big\rangle.$$

- For $\Omega \subset \mathbb{R}^n$ open, $f, g : \Omega \to \mathbb{R}$ partially differentiable, $\lambda \in \mathbb{R}$ and $x \in \Omega$ we have

$$\nabla(f+g)(x) = \nabla f(x) + \nabla g(x), \qquad \nabla(\lambda f)(x) = \lambda \nabla f(x),$$

$$\nabla(fg)(x) = g(x)\nabla f(x) + f(x)\nabla g(x).$$

  The first two equations imply:

$$\nabla : \{f : \Omega \to \mathbb{R} :\ f \text{ partially differentiable}\} \to \{f : \Omega \to \mathbb{R}^n\}$$

  is a linear mapping (or: linear operator, see functional analysis!).

Example: $r : \mathbb{R}^n \to \mathbb{R}$, $r(x) = \|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$. For $x \in \mathbb{R}^n$, $x \neq 0$ we have

$$\partial_i r(x) = \frac{1}{2}\left(\sum_{i=1}^n x_i^2\right)^{-\frac{1}{2}} \cdot 2x_i = \frac{x_i}{\|x\|_2},$$

and thus

$$\nabla r(x) = \frac{x}{\|x\|_2}.$$

Definition (Laplacian)**.** *Let $\Omega \subset \mathbb{R}^n$ be open, $f \in C^2(\Omega)$. Then for $x \in \Omega$*

$$\Delta f(x) := \sum_{i=1}^n \partial_i^2 f(x)$$

is called the ***Laplacian*** *of f at x. $\Delta : C^2(\Omega) \to C^0(\Omega)$ is called the Laplacian operator.*

Examples:

- Let $\Omega \subset \mathbb{R}^n$ open, $I \subset \mathbb{R}$ open, $f : \Omega \times I \to \mathbb{R}$, $k \in \mathbb{R}$. Then

$$\partial_t f(x,t) - k\Delta f(x,t) = 0$$

  is called the heat equation.

- Let $x_0 \in \mathbb{R}^3$, $N, Z \in \mathbb{N}$, then

$$H_{\text{atom}} = -\sum_{i=1}^{N} \Delta_{x_i} + \sum_{i=1}^{N} \frac{-Z}{|x_i - x_0|} + \sum_{1 \le i < j \le N} \frac{1}{|x_i - x_j|}$$

  with $H_{\text{atom}} : L^2(\mathbb{R}^{3N}) \to L^2(\mathbb{R}^{3N})$, $\mathcal{D}(H_{\text{atom}}) = H^2(\mathbb{R}^{3N})$ is called the Schrödinger operator for atoms *(see quantum mechanics, spectral theory).*

Definition (Hessian)**.** *Let $\Omega \subset \mathbb{R}^n$ open, $f \in C^2(\Omega)$. The matrix*

$$H_f(x) = \begin{pmatrix} \partial_1^2 f(x) & \cdots & \partial_n \partial_1 f(x) \\ \vdots & \ddots & \vdots \\ \partial_1 \partial_n f(x) & \cdots & \partial_n^2 f(x) \end{pmatrix}$$

*is called the **Hessian matrix**, or **Hessian** of f in $x \in \Omega$.*

Property: $H_f(x)$ is symmetric!

Example: $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^2$. Then $\nabla f(x, y) = (2x, 2y)^T$ and $H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$.

2.5.2 Total derivatives

Suppose $f : \mathbb{R} \to \mathbb{R}$, $f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$. Equivalently we have

$$f(x + h) - f(x) = f'(x)h + \text{rest}(h), \qquad \lim_{h \to 0} \frac{\text{rest}(h)}{h} = 0,$$

that is the growth $f(x + h) - f(x)$ is approximated by the linear mapping $h \mapsto f'(x)h$.

Now suppose $f : \mathbb{R}^n \to \mathbb{R}^m$, $f(x) = \big(f_1(x), \ldots, f_m(x)\big)^T$, we can look at $\partial_i f_j$, but a general "slope" concept like $f'(x)$ is not applicable. But we can approximate using a linear mapping!

Definition (Total derivative)**.** *Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}^m$. f is called **differentiable** in $x \in \Omega$, if there exists a linear mapping $T : \mathbb{R}^n \to \mathbb{R}^m$ such that*

$$\lim_{h \to 0} \frac{\|f(x + h) - f(x) - T(h)\|}{\|h\|} = 0.$$

*The mapping T is called the **derivative** of f in x, or the **total derivative** or the **Fréchet derivative**. It is also denoted as $Df(x)$.*

Remarks:

- Thanks to the equivalence of norms on $\mathbb{R}^n/\mathbb{R}^m$, the employed norms do not matter.
- If the total derivative exists, the mapping T is unique.

Example: $f : \mathbb{R}^n \to \mathbb{R}^m$, $f(x) = Ax + b$ for $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. Then for $h \in \mathbb{R}^n$

$$f(x + h) - f(x) - Ah = A(x + h) + b - (Ax + b) - Ah = 0,$$

thus $Df(x) = A$, independent of $x \in \mathbb{R}^n$.

Theorem (Jacobi matrix)**.** *Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}^m$ with $f(x) = \big(f_1(x), \dots, f_m(x)\big)^T$. If f is totally differentiable in $x \in \Omega$, then the components $f_1, \dots, f_m$ are partially differentiable in x, f is continuous in x and we have*

$$Df(x) = J_f(x) := \begin{pmatrix} \partial_1 f_1(x) & \cdots & \partial_n f_1(x) \\ \vdots & \ddots & \vdots \\ \partial_1 f_m(x) & \cdots & \partial_n f_m(x) \end{pmatrix}.$$

*$J_f(x)$ is called the **Jacobi matrix** of f in x, or the **Jacobian**.*

Conversely, if all components $f_1, \dots, f_m : \Omega \to \mathbb{R}$ of $f : \Omega \to \mathbb{R}^m$ are continuously differentiable in $x \in \Omega$, then f is totally differentiable in x with $Df(x) = J_f(x)$.

Remark: We have for $f : \Omega \to \mathbb{R}^m$, $\Omega \subset \mathbb{R}^n$ open:

$$\begin{aligned} f \text{ continuously differentiable } \textit{(i.e. partial diff.)} &\overset{(*)}{\Longrightarrow} f \text{ differentiable } \textit{(i.e. total diff.)} \\ f \text{ differentiable } \textit{(i.e. total diff.)} &\Longrightarrow f \text{ partially differentiable} \\ f \text{ differentiable } \textit{(i.e. total diff.)} &\Longrightarrow f \text{ continuous} \end{aligned}$$

Any other implication between these four properties of f is false!

Example: Transformation from polar coordinates in the plane to cartesian coordinates:

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \qquad f(r, \phi) = \big(r\cos(\phi), r\sin(\phi)\big).$$

We can calculate

$$J_f(r, \phi) = \begin{pmatrix} \cos(\phi) & -r\sin(\phi) \\ \sin(\phi) & r\cos(\phi) \end{pmatrix}$$

proving that f is differentiable (totally) with $Df = J_f$, see $(*)$. J_f with $r = 1$ should seem familiar!

Theorem (Chain rule)**.** *Let $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ open and let $f : U \to V$, $g : V \to \mathbb{R}^l$. If f is differentiable in $x \in U$ and g differentiable in $f(x) \in V$, then $g \circ f$ is differentiable in x and*

$$\begin{aligned} D(g \circ f)(x) &= Dg\big(f(x)\big) \circ Df(x) \\ J_{g\circ f}(x) &= J_g\big(f(x)\big) \cdot J_f(x). \end{aligned}$$

*This is the **multi–dimensional chain rule**.*

Remark: You can rewrite $J_{g\circ f}(x) = J_g\big(f(x)\big) \cdot J_f(x)$ as

$$\partial_j (g \circ f)_i(x) = \sum_{k=1}^{m} (\partial_k g_i)\big(f(x)\big) \cdot \partial_j f_k(x)$$

Example: Let $f : \mathbb{R} \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$ differentiable. We can calculate $J_{g\circ f}$ for $x \in \mathbb{R}$ as:

$$(g \circ f)'(x) = \sum_{k=1}^{m} (\partial_k g)\big(f(x)\big) \cdot f_k'(x) = \big\langle \nabla g(f(x)), f'(x) \big\rangle,$$

where $f'(x) := Df(x)$.

2.5.3 Subdifferentials

Some important functions have bends at which they are not differentiable. One prominent example is the absolute value function $f : \mathbb{R} \to \mathbb{R}$, $f(z) = |z|$, which has a slope of -1 for $z < 0$ and a slope of $+1$ for $z > 0$, and is thus not differentiable at $z = 0$. For these cases, it is sometimes useful to introduce a less stringent notion of derivative:

Definition (Subgradient, Subdifferential)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. A vector $g \in \mathbb{R}^n$ is called **subgradient** of f at $x \in \mathbb{R}^n$, if for all $z \in \mathbb{R}^n$ we have*

$$f(z) \geq f(x) + g^T(z - x).$$

*We call f **subdifferentiable** at $x \in \mathbb{R}^n$ if there exists at least one subgradient at x. The set of subgradients of f at x is called the **subdifferential** of f at x, which is often denoted as $\partial f(x)$. The function f is called **subdifferentiable**, if it is subdifferentiable for all $x \in \mathbb{R}^n$.*

Example: Let $f : \mathbb{R} \to \mathbb{R}$, $f(z) = |z|$, the absolute value function. For $x < 0$, the subgradient is unique, namely $\partial f(x) = \{-1\}$. For $x > 0$, we similarly have $\partial f(x) = \{+1\}$. For $x = 0$, the subgradient inequality states $|z| \geq gz$, which is fulfilled in case of $g \in [-1, +1]$. Hence we have $\partial f(0) = [-1, +1]$.

Properties: Let $f : \mathbb{R}^n \to \mathbb{R}$.

- The subdifferential $\partial f(x)$ is always a closed, convex set (see section 3 for more details on convexity).
- All convex and continuous functions are subdifferentiable.
- If f is convex and differentiable in x, then $\partial f(x) = \{\nabla f(x)\}$.
- If f is convex and $\partial f(x) = \{g\}$, then f is differentiable in x and $\nabla f(x) = g$.

We will be using subgradients later for optimization, see section 3.

2.6 Taylor expansion

Definition (Taylor series / expansion)**.** *Let $f \in C^\infty(\mathbb{R})$. If f is sufficiently "nice", then for $h \in \mathbb{R}$*

$$f(x + h) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x)}{k!} h^k$$

*is called **Taylor series** of f around x.*

Even if f is not "nice", we have for arbitrary $l \in \mathbb{N}$

$$f(x + h) = \sum_{k=0}^{l} \frac{f^{(k)}(x)}{k!} h^k + rest(h)$$

*with $\lim_{h \to 0} rest(h) = 0$. This is called the **Taylor expansion** of f around x.*

Example: Calculate approximation to special functions, for example sin, cos, etc. We know $\sin'(x) = \cos(x)$, $\sin''(x) = -\sin(x)$, $\sin'''(x) = -\cos(x)$, ..., and that $\sin(0) = 0$, $\cos(0) = 1$. Using Taylor expansion we then have

$$\begin{aligned}\sin(0+x) &= \frac{\sin(0)}{0!}x^0 + \frac{\cos(0)}{1!}x^1 + \frac{-\sin(0)}{2!}x^2 + \frac{-\cos(0)}{3!}x^3 + \ldots + \text{rest}(x) \\ &= 0 + x + 0 - \frac{x^3}{3!} + \ldots + \text{rest}(x).\end{aligned}$$

A possible approximation to sin is thus for example

$$\sin(x) \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}.$$

In fact we can prove

$$\sin(x) = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)!} x^{2j+1},$$

once a proper notion of convergence for infinite sums has been defined.

If you want to try out this example or others (like $\exp(x)$, $\log(1+x)$) in Matlab, the function `taylortool` is your friend.

Definition (Multi–index)**.** *Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a **multi–index** with operations $|\alpha| := \sum_{i=1}^n \alpha_i$, $\alpha! := \prod_{i=1}^n \alpha_i!$ and $x^\alpha := \prod_{i=1}^n x_i^{\alpha_i}$ for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. For $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^{|\alpha|}(\mathbb{R}^n)$, define*

$$\partial^\alpha f := \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_n^{\alpha_n} f.$$

Definition (Multi–dimensional Taylor expansion)**.** *Let $f \in C^l(\mathbb{R}^n)$ for $l \in \mathbb{N}$, α a multi–index. The **Taylor expansion** of f around $x \in \mathbb{R}^n$ is*

$$f(x+h) = \sum_{|\alpha| \leq l} \frac{\partial^\alpha f(x)}{\alpha!} h^\alpha + \mathit{rest}(h)$$

for $h \in \mathbb{R}^n$ with $\lim_{h \to 0} \frac{\mathit{rest}(h)}{\|h\|^l} = 0$.

Example: Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}$ a scalar field with $f \in C^2(\Omega)$. The often used second–order Taylor expansion of f around x is

$$f(x+h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T H_f(x) h + \text{rest}(h)$$

for $x + h \in \Omega$, with $\lim_{h \to 0} \frac{\text{rest}(h)}{\|h\|^2} = 0$.

2.7 Case Study (Part I)

(adapted from Maximilian Baust)

Suppose we are given a video sequence like the one depicted in Fig. 1 and we want to track some objects (in this case: cars) in them. For this reason we make the following assumptions about the changes from one image to another:

Figure 1: Three pictures of a video sequence by Dirk-Jan Kroon (MATLAB central files exchange, Lucas Kanade affine template tracking)

- All changes are very small (in comparison to the image size).
- All changes can be described by a pure translation.

We will see that these assumptions may be too simplifying, but for educational purposes it is sufficient to start with them. Moreover, we will not try to track whole objects, but rather points on them.

Before we can model the tracking itself, we have to model an image. In (computer vision) theory an image is often modeled as a continuous function $I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$. In practice we have, of course, only a discretized version of it, which means that we consider the pixels of the discrete image as measurements of I. However, it is often better to develop the theoretical framework with the idealized image I.

In order to track a point in our video sequence we will consider what happens at one point between two subsequent images I_s (source image) and I_t. Obviously the intensity value of this point will be transported by a vector $v \in \mathbb{R}^2$ such that it will be identical with an intensity value in the target image I_t. Since we assumed that all movements can be approximated with a translation we can model this by

$$I_s(x + v) = I_t(x). \tag{1}$$

Unfortunately v has two entries and thus one equation is not enough for computing v. For this reason we simply take more measurements, in our case pixels, into account and assume that v is approximately constant for a whole patch $P \subset \Omega$. This leads us to the following minimization problem:

$$\min_{v \in \mathbb{R}^2} \int_P |I_s(x + v) - I_t(x)|^2 \, dx. \tag{2}$$

The only thing that is disturbing now, is the fact that v is an argument of I_s, a nonlinear problem. To solve this problem, we approximate $I_s(x + v)$ with a Taylor expansion (a linearization)

$$I_s(x + v) \approx I_s(x) + \nabla I_s(x)^T v.$$

Instead of solving (2) we now solve

$$\min_{v \in \mathbb{R}^2} \int_P |I_s(x) + \nabla I_s(x)^T v - I_t(x)|^2 \, dx. \tag{3}$$

I_s I_t

Figure 2: We assume that all changes occurring in the red patches of the source image I_s and the target image I_t can be described by the same translation vector v.

In discrete formulation this reads as

$$\min_{v\in\mathbb{R}^2} \sum_{x\in P_h} \left|\nabla I_s(x)^T v + I_s(x) - I_t(x)\right|^2 \tag{4}$$

where P_h is a discrete patch with $x \in P_h$ a discrete pixel.

Let now $P_h = \{x_1, \ldots, x_n\}$ and let

$$A = \begin{pmatrix} \partial_1 I_s(x_1) & \partial_2 I_s(x_1) \\ \vdots & \vdots \\ \partial_1 I_s(x_n) & \partial_2 I_s(x_n) \end{pmatrix}, \qquad b = \begin{pmatrix} I_t(x_1) - I_s(x_1) \\ \vdots \\ I_t(x_n) - I_s(x_n) \end{pmatrix}$$

then we can formulate our problem (4) as

$$\min_{v\in\mathbb{R}^2} \sum_{x\in P_h} \left|\nabla I_s(x)^T v + I_s(x) - I_t(x)\right|^2 = \min_{v\in\mathbb{R}^2} \|Av - b\|_2^2,$$

a least squares problem! Now we can solve this problem from frame to frame for example by using the algorithm of Golub.

3 Optimization

Motivation: Recall the series expansion method from chapter 1: we want to reconstruct a signal $f : V \to \mathbb{R}$ for some $V \subset \mathbb{R}^3$. We expand $f \approx \hat{f} = \sum_i x_i b_i$ using basis functions b_i. Using measurements (m_j) and a measurement model $\mathcal{M}_j : f \mapsto m_j$. Then

$$m_j = \sum_i x_i \, \mathcal{M}_j b_i.$$

Using shorthand $a_{ji} := \mathcal{M}_j b_i$, $A := (a_{ji})$, $m := (m_j)$ and $x := (x_i)$ we receive the linear system equation

$$Ax = m.$$

This linear system is inconsistent due to measurement and model errors, hence we try to solve a least squares problem instead,

$$\min_x \|Ax - m\|_2^2.$$

This is a typical optimization problem.

General setting: Let $(V, \|\cdot\|)$ be a normed vector space, $f : V \to \mathbb{R}$. Consider the problem

$$\min_{x \in V} f(x).$$

(For a maximization problem consider $-f$ instead.) We are interested in:

- existence of minimizer, or: is there a $x^* \in V$ with $f(x^*) = \min_{x \in V} f(x)$? $\to$ 3.1
- uniqueness of minimizer, or: if we have x^* minimizer, is it the only one? $\to$ 3.2
- is some $x \in V$ a local minimizer? $\to$ 3.3
- an algorithm or method to find a local minimizer $\to$ 3.4 onwards

3.1 Existence of a minimizer

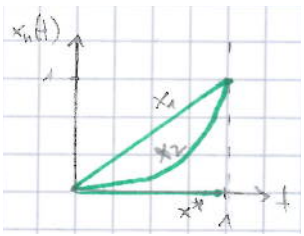
Example: $V = \big\{x \in C([0,1]) : \ x : [0,1] \to \mathbb{R}$ continuous with $x(0) = 0$, $x(1) = 1\big\}$.

Let $f : V \to \mathbb{R}$ be defined by

$$f(x) := \int_0^1 |x(t)|^2 dt.$$

Problem: $\min_{x \in V} f(x)$.

We have $f \geq 0$. A trivial minimizer would be $x^* \equiv 0$ with $f(x^*) = 0$, however $x^* \notin V$. What about $x_n : [0,1] \to \mathbb{R}$, $x_n(t) := t^n$ for $n \in \mathbb{N}$? We have $\lim_{n\to\infty} f(x_n) = 0$, but $x^*(t) = \lim_{n\to\infty} x_n(t) = \begin{cases} 1 & t = 1 \\ 0 & t \neq 1 \end{cases}$ and thus $x^* \notin V$.



Example: $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \cos(x)$. *Problem:* $\min_{x\in\mathbb{R}} f(x)$.
We have $f(x) \geq -1$. Define $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ as

$$x_n := \begin{cases} -\pi + \frac{1}{n} & \text{for } n \text{ even} \\ \pi & \text{for } n \text{ odd,} \end{cases}$$

then $\lim_{n\to\infty} f(x_n) = -1$, but (x_n) is not convergent. A subsequence, however, would do the job. This works in general for compact sets, here we have $(x_n) \subset [-\pi, \pi]$, and $[-\pi, \pi]$ is compact in $\mathbb{R}$.

Main Example: Let (M, d) be a metric space, $f : M \to \mathbb{R}$. Assume that f is bounded from below, that is $\exists c \in \mathbb{R}$ such that $f(x) \geq c\ \forall x \in M$. Furthermore, assume there exists a minimizing sequence $(x_k)_{k\in\mathbb{N}} \subset M$ with $\lim_{k\to\infty} f(x_k) = c$.

- If M is compact, there exists a convergent subsequence $(x_{k_j})_{j\in\mathbb{N}}$ with $\lim_{j\to\infty} x_{k_j} = x^*$.
- If f is continuous, then $f(M)$ is compact, hence closed and $f(x^*) = c = \min_{x\in M} f(x)$.

Theorem (Existence of minimizer)**.** *Let (M, d) be a metric space, $U \subset M$ and $f : U \to \mathbb{R}$. If U is compact and f continuous, then there exists $x^* \in U$ such that*

$$f(x^*) = \min_{x\in U} f(x).$$

3.2 Uniqueness of a minimizer / Convexity

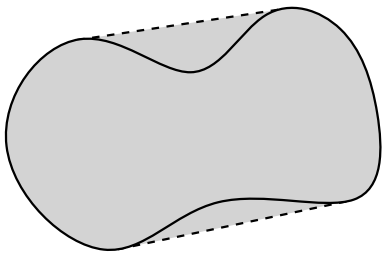
Definition (Convex set / convex hull)**.** *Let V be a linear space, $A \subset V$. A is called **convex**, if*

$$\lambda x + (1 - \lambda)y \in A$$

for all $x, y \in A$ and $\lambda \in [0, 1]$.
The set

$$conv(A) := \left\{ \sum_{i=1}^{k} \lambda_i x_i : \ x_i \in A,\ \lambda_i \in [0, 1] \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

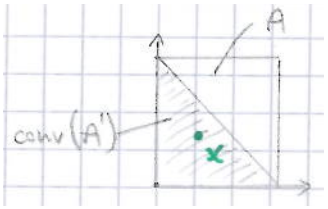
*is called **convex hull** of A, the set of all convex combinations.*



Theorem. *Let V be a vector space, $A \subset V$.*

$$A \text{ convex} \qquad \Longleftrightarrow \qquad conv(A) = A.$$

Theorem (Carathéodory)**.** *Let $A \subset \mathbb{R}^n$ and $x \in conv(A)$. Then there exists $A' \subset A$, $|A'| \leq n+1$, such that $x \in conv(A')$.*



Definition (Convex function)**.** *Let V be a linear space, $A \subset V$ convex. A function $f : A \to \mathbb{R}$ is called **convex**, if*

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$

for all $x, y \in A$ and $\lambda \in [0,1]$.

*f is called **strictly convex** if*

$$f\big(\lambda x + (1-\lambda)y\big) < \lambda f(x) + (1-\lambda) f(y)$$

for all $x, y \in A$, $x \neq y$ and $\lambda \in (0,1)$.

Example: $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_p$ for $p \geq 1$ is convex.

(a) $p = 1$ (b) $p = 2$ (c) $p = \infty$

Theorem. *Let $\Omega \subset \mathbb{R}^n$ be open, $f \in C^2(\Omega)$. Then*

$$f \text{ convex} \quad \Longleftrightarrow \quad H_f(x) \text{ positive semi-definite } \forall x \in \Omega.$$

and

$$H_f(x) \text{ positive definite } \forall x \in \Omega \quad \Longrightarrow \quad f \text{ strictly convex.}$$

Theorem (Uniqueness of minimizer)**.** *Let V be a linear space, $A \subset V$ convex, $f : A \to \mathbb{R}$ strictly convex. If f has a minimizer x^* with $f(x^*) = \min_{x \in A} f(x)$, then x^* is unique.*

Proof: Assume there exists another minimizer $\bar{x}$ with $f(\bar{x}) = f(x^*) = c$. As f is strictly convex we have for $\lambda \in (0,1)$

$$f\big(\lambda x^* + (1-\lambda)\bar{x}\big) < \lambda f(x^*) + (1-\lambda) f(\bar{x}) = c,$$

a contradiction. □

3.3 Identifying local minima

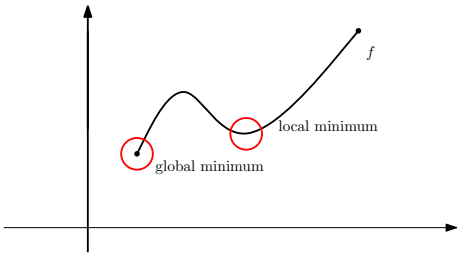
Definition (Local / global minimum / maximum)**.** *Let $\Omega \subset \mathbb{R}^n$, $f : \Omega \to \mathbb{R}$. $x \in \Omega$ is called **local minimum / local maximum** of f, if there exists a neighborhood U of x such that*

$$f(x) \leq f(y) \qquad \text{respectively} \qquad f(x) \geq f(y) \qquad \text{for all } y \in \Omega \cap U.$$

*In both cases x is called **local extremal point**.*

*A local extremal point is called **isolated** if additionally we have $f(y) \neq f(x)$ for all $y \in \Omega \cap U$ with $y \neq x$.*

*If $U = \Omega$, then x is called a **global minimum / global maximum** or **global extremum**.*



Theorem. *Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}$ partially differentiable. If $x \in \Omega$ is an extremal point of f, then $\nabla f(x) = 0$.*

Theorem. *Let $\Omega \subset \mathbb{R}^n$ open, $f \in C^2(\Omega)$, $x \in \Omega$. Then*

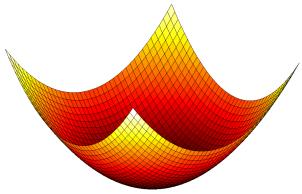
$$\begin{aligned} x \text{ local minimum} &\implies H_f(x) \text{ positive semi–definite} \\ x \text{ local maximum} &\implies H_f(x) \text{ negative semi–definite} \end{aligned}$$

Conversely,

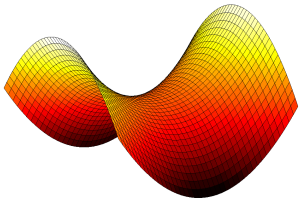
$$\begin{aligned} \nabla f(x) = 0,\ H_f(x) \text{ positive definite} &\implies x \text{ isolated local minimum} \\ \nabla f(x) = 0,\ H_f(x) \text{ negative definite} &\implies x \text{ isolated local maximum} \end{aligned}$$

Examples:

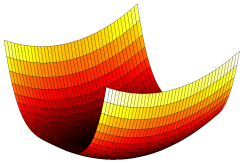
- $f_1 : \mathbb{R}^2 \to \mathbb{R}$, $f_1(x, y) = x^2 + y^2$. $\nabla f_1(x, y) = (2x, 2y)^T$, $H_{f_1}(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. We have $\nabla f_1(0) = 0$, $H_{f_1}(0)$ positive definite, thus 0 is an isolated local minimum.
- $f_2 : \mathbb{R}^2 \to \mathbb{R}$, $f_2(x, y) = x^2 - y^2$. $\nabla f_2(x, y) = (2x, -2y)^T$, $H_{f_2}(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$. We have $\nabla f_2(0) = 0$, but $H_{f_2}(0)$ is indefinite.
- $f, g, h : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 + y^4$, $g(x, y) = x^2$, $h(x, y) = x^2 + y^3$. Then $\nabla f(0) = \nabla g(0) = \nabla h(0) = 0$ and $H_f(0) = H_g(0) = H_h(0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ positive semi–definite. No conclusion is possible. In fact, f has an isolated minimum at 0, g has a non–isolated local minimum at 0 and h has no local extremum at all.



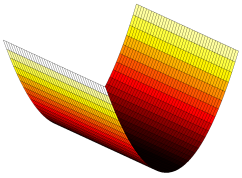
(a) $f_1(x, y) = x^2 + y^2$



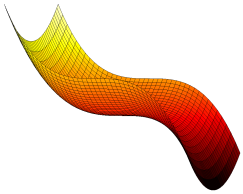
(b) $f_2(x, y) = x^2 - y^2$



(c) $f(x, y) = x^2 + y^4$



(d) $g(x, y) = x^2$



(e) $h(x, y) = x^2 + y^3$

Figure 3: Graphs of the functions f_1, f_2 and f, g, h from the example about local extrema.

Let $\Omega \subset \mathbb{R}^n$ open, $f : \Omega \to \mathbb{R}$, $f \in C^2(\Omega)$. The main result is:

Necessary condition for a local minimum: If $x^* \in \Omega$ is a local minimizer of f, then

$$\nabla f(x^*) = 0.$$

Sufficient condition for a local minimum: Let $x^* \in \Omega$. If

$$\nabla f(x^*) = 0 \qquad \text{and} \qquad H_f(x^*) > 0,$$

then f has a local minimum at x^*.

Proof (informal): First, x^* is a candidate since $\nabla f(x^*) = 0$. Second, $f \in C^2(\Omega)$, so $H_f(x)$ is continuous. Thus there exists a bounded subset $U \subset \Omega$ with $x^* \in U$ and $H_f(x) > 0$ for all $x \in U$. Third, assume without loss of generality $U = \overline{U}$, so we have a continuous f on a compact $U \overset{3.1}{\Longrightarrow}$ there exists a minimizer $\bar{x}$ in U. Fourth, f is strictly convex in U, thus $\bar{x}$ is the only minimizer and $x^* = \bar{x}$.

3.4 Gradient descent

For this section we want to minimize a function $f : \mathbb{R}^n \to \mathbb{R}$:

$$\min_{x \in \mathbb{R}^n} f(x).$$

We try to solve this problem iteratively. We start at an initial value $x_0 \in \mathbb{R}^n$ and iteratively compute updates

$$x_{k+1} = x_k + \alpha_k d_k,$$

where $d_k \in \mathbb{R}^n$ is a **direction of improvement** and $\alpha_k > 0$ is the **step size**.

Definition (Direction of Improvement)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. A vector $d \in \mathbb{R}^n$ is called* ***direction of improvement****, if there exists $\alpha > 0$ such that*

$$f(x + \alpha d) < f(x).$$

Property: If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, every $d \in \mathbb{R}^n$ with $\nabla f(x)^T d < 0$ is a direction of improvement.

Proof: $\alpha \mapsto f(x + \alpha d)$ is a function $(0, \infty) \to \mathbb{R}$, where

$$\frac{\partial}{\partial \alpha} f(x + \alpha d) = \nabla f(x + \alpha d)^T d.$$

Since $\nabla f(x)^T d < 0$, f decreases in this direction. □

A natural choice is thus $d_k = -\nabla f(x_k)$, yielding:

Gradient descent. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, $x_0 \in \mathbb{R}^n$ an initial value. A minimizing sequence $(x_n)_{n \in \mathbb{N}_0} \subset \mathbb{R}^n$ with $f(x_0) \geq f(x_1) \geq f(x_2) \geq \ldots$ can be constructed by

$$x_{k+1} = x_k + \alpha_k d_k,$$

where $d_k = -\nabla f(x_k)$ and $\alpha_k > 0$ is the step size corresponding to d_k. α_k can be computed as $\alpha_k = \min_{\alpha > 0} f(x_k + \alpha d_k)$.

Brief excursion: subgradient method. If our function $f : \mathbb{R}^n \to \mathbb{R}$ is not differentiable, we might still have the *subgradient* available. Using it, we can define a method similar to gradient descent:

Let $f : \mathbb{R}^n \to \mathbb{R}$ convex and continuous, and $x_0 \in \mathbb{R}^n$ an initial value. The sequence

$$x_{k+1} = x_k - \alpha_k g_k,$$

where g_k is any subgradient of f at x_k, and $\alpha_k \geq 0$ are pre-determined constants, is called the **subgradient method.**

Please note, that the step size α_k has to be constant and has to be chosen before the method runs (in contrast to gradient descent). For proper choices of the constants, the subgradient method will converge (but slowly).

Also note that the subgradient method is not a descent method, that is, a step may actually increase the function value! A common work-around is to set

$$f_{k+1}^{best} = \min\left(f_k^{best}, f(x_k)\right),$$

and to record which iteration actually yields the best function value of f.

Gradient descent for quadratic forms. Let's assume we want to study the following optimization problem based on a quadratic form f: Let $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n \times n}$ symmetric positive definite, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$ and consider

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{with} \quad f(x) = \tfrac{1}{2} x^T A x - x^T b + c.$$

We already know that

$$\nabla f(x) = \tfrac{1}{2} A^T x + \tfrac{1}{2} A x - b = Ax - b,$$

and $H_f(x) = A$. Since A is also positive definite, this means a solution x^* to $Ax - b = 0$ (or $Ax = b$) is the unique minimizer of f.

Assuming inversion of A is not feasible (for example because of problem size or condition of A), we now apply gradient descent. It turns out that for this type of f, we can compute α_k by noting

$$\begin{aligned} 0 &\overset{!}{=} \frac{\partial}{\partial \alpha_k} f(x_k + \alpha_k d_k) = \nabla f(x_{k+1})^T d_k = (Ax_{k+1} - b)^T d_k \\ &= \underbrace{(Ax_k - b)^T}_{=\nabla f(x_k)} d_k + \alpha_k (Ad_k)^T d_k = -d_k^T d_k + \alpha_k d_k^T A d_k, \end{aligned}$$

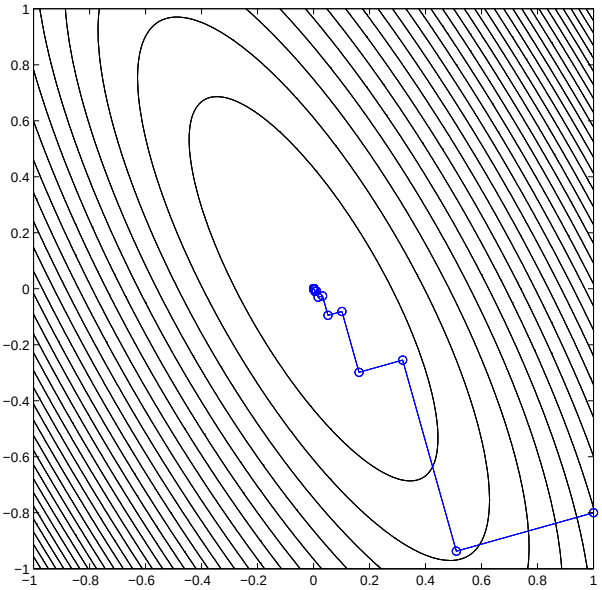
where we used $x_{k+1} = x_k + \alpha_k d_k$ and $d_k = -\nabla f(x_k)$, to arrive at

$$\alpha_k = \frac{d_k^T d_k}{d_k^T A d_k}.$$

Zig–Zagging. Using for example the quadratic form f with

$$A = \begin{pmatrix} 7.75 & 3.9 \\ 3.9 & 3.25 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{and} \quad c = 0,$$

the gradient descent method starting at $x_0 = \left(\begin{smallmatrix} 1 \\ -0.8 \end{smallmatrix}\right)$ shows very bad convergence:



The gradient is always orthogonal to the isolines and shows the characteristic *zig-zagging* in this example. A better method using a more appropriate concept of orthogonality is the conjugate gradient method.

3.5 Conjugate gradient

We again study the optimization problem from the last section: $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n\times n}$ symmetric positive definite, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$ with

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{with} \quad f(x) = \tfrac{1}{2}x^T Ax - x^T b + c.$$

Recall for $x, y \in \mathbb{R}^n$ the *energy scalar product*

$$\langle x, y\rangle_A = \langle x, Ay\rangle = x^T Ay$$

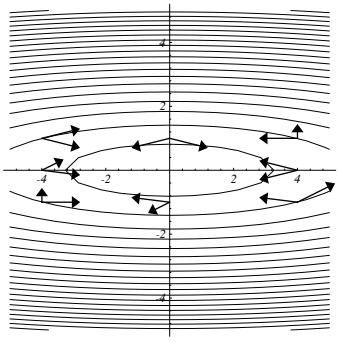
and the induced *energy norm*

$$\|x\|_A = \sqrt{\langle x, x\rangle_A}.$$

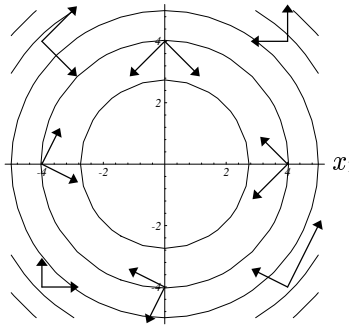
Definition (Conjugate vectors)**.** *Let $A \in \mathbb{R}^{n\times n}$ symmetric positive definite. Two vectors $x, y \in \mathbb{R}^n$ are called A–**orthogonal** or **conjugate** if*

$$\langle x, y\rangle_A = 0.$$

A picture from Jonathan Shewchuk illustrates the modified concept of orthogonality:



These pairs of vectors are A-orthogonal ...



... because these pairs of vectors are orthogonal

The idea of the CG method is now to use the gradient descent scheme, i.e. choose $x_0 \in \mathbb{R}^n$ and then iterate using directions of improvement d_k and step sizes α_k

$$x_{k+1} := x_k + \alpha_k d_k, \tag{5}$$

where the d_k are chosen to be A–orthogonal, that is $\langle d_i, d_j\rangle_A = 0$ for $i \neq j$. The step size α_k is computed as previously as $\alpha_k = \min_{\alpha>0} f(x_k + \alpha d_k)$.

We introduce the **residual**

$$r_k := b - Ax_k$$

and the **error**

$$e_k := x_k - x^*,$$

where x^* is the solution to our minimization problem $\min_{x\in\mathbb{R}^n} f(x)$. We have

$$r_k = b - Ax_k = Ax^* - Ax_k = -A(x_k - x^*) = -Ae_k \tag{6}$$

and

$$e_{k+1} = x_k + \alpha_k d_k - x^* = e_k + \alpha_k d_k \tag{7}$$

Thus our α_k can be computed as

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\partial}{\partial \alpha_k} f(x_k + \alpha_k d_k) = \nabla f(x_{k+1})^T d_k = -r_{k+1}^T d_k = d_k^T A e_{k+1} \\ &= d_k^T A(e_k + \alpha_k d_k) = d_k^T A e_k + \alpha_k d_k^T A d_k, \end{aligned}$$

which yields

$$\alpha_k = -\frac{d_k^T A e_k}{d_k^T A d_k} = \frac{d_k^T r_k}{d_k^T A d_k}. \tag{8}$$

This also implies $\langle e_{k+1}, d_k \rangle_A = 0$, that is the error e_{k+1} is conjugate to the direction of improvement d_k.

Proposition. The iteration scheme (5) computes the solution x^* to $\min_{x \in \mathbb{R}^n} f(x)$ in n steps using conjugate directions d_k and α_k as computed in (8).

Proof: We can write e_0 as a linear combination of the d_k,

$$e_0 = \sum_{j=0}^{n-1} \delta_j d_j.$$

To compute the δ_j we multiply by $d_k^T A$ from the left and use the conjugacy of the d_k,

$$d_k^T A e_0 = \sum_{j=0}^{n-1} \delta_j d_k^T A d_j = \delta_k d_k^T A d_k,$$

which yields

$$\delta_k = \frac{d_k^T A e_0}{d_k^T A d_k} = \frac{d_k^T A \left(e_0 + \sum_{j=0}^{k-1} \alpha_j d_j \right)}{d_k^T A d_k} = \frac{d_k^T A e_k}{d_k^T A d_k} = -\alpha_k.$$

Thus we have

$$e_k = e_0 + \sum_{j=0}^{k-1} \alpha_j d_j = \sum_{j=0}^{n-1} \delta_j d_j - \sum_{j=0}^{k-1} \delta_j d_j = \sum_{j=k}^{n-1} \delta_j d_j, \tag{9}$$

which implies $e_n = 0$. □

Conjugate directions

Now the open question is how to choose the conjugate d_k. For this we choose any basis of $\mathbb{R}^n$, for example $u_0, \ldots, u_{n-1}$ and force it to be conjugate, that is we remove the parts that are not A–orthogonal by setting:

$$\begin{aligned} d_0 &:= u_0 \\ d_k &:= u_k + \sum_{j=0}^{k-1} \beta_{kj} d_j \qquad \text{for } k = 1, \ldots, n-1. \end{aligned} \tag{10}$$

The $\beta_{kj} \in \mathbb{R}$ are only defined for $k > j$ and can be computed similar to the δ_j earlier:

$$d_k^T A d_j = u_k^T A d_j + \sum_{i=0}^{k-1} \beta_{ki} d_i^T A d_j$$

and thus for $k > j$ due to conjugacy of the d_k

$$0 = u_k^T A d_j + \beta_{kj} d_j^T A d_j$$

yielding

$$\beta_{kj} = -\frac{u_k^T A d_j}{d_j^T A d_j}. \tag{11}$$

This is the **method of conjugate directions**, it converges in n steps, but it requires to keep all search directions in memory and has relatively high complexity ($O(n^3)$). If you choose the canonical basis $e_1, \ldots, e_n$, this method corresponds to the traditional Gaussian elimination.

This method generates a sequence of subspaces

$$\mathcal{D}_k := \operatorname{span}\{d_0, \ldots, d_{k-1}\}$$

with $\mathcal{D}_0 \subset \mathcal{D}_1 \subset \ldots \subset \mathcal{D}_{n-1}$ such that each error e_k lies in the affine subspace $e_0 + \mathcal{D}_k$ with $\|e_k\|_A = \min$ for each k. The resulting x_k is also called a **Ritz–Galerkin approximation**.

Conjugate gradient

Now we can finally formulate the CG method: choose as $u_k = r_k = b - Ax_k$, just as we did for gradient descent. This allows us to compute the β_{kj} in a more optimal fashion.

First we note that $r_k \perp \mathcal{D}_k$, as we have due to (6) and (9)

$$d_i^T r_k = -d_i^T A e_k = -\sum_{j=k}^{n-1} \delta_j d_i^T A d_j$$

and thus for $i < k$

$$d_i^T r_k = 0. \tag{12}$$

Furthermore we have

$$d_k^T r_k = r_k^T r_k \tag{13}$$

as per construction (10)

$$d_k^T r_k = r_k^T r_k + \underbrace{\sum_{j=0}^{k-1} \beta_{kj} d_j^T r_k}_{=0}.$$

Second, we have per construction

$$\mathcal{D}_k = \operatorname{span}\{r_0, \ldots, r_{k-1}\}.$$

Using

$$r_{k+1} = -A e_{k+1} = -A(e_k + \alpha_k d_k) = r_k - \alpha_k A d_k \tag{14}$$

we also have

$$\begin{aligned}\mathcal{D}_k &= \operatorname{span}\{d_0, Ad_0, A^2d_0, \ldots, A^{k-1}d_0\}\\ \mathcal{D}_k &= \operatorname{span}\{r_0, Ar_0, A^2r_0, \ldots, A^{k-1}r_0\}\end{aligned}$$

This kind of subspace is also called **Krylov subspace**. As we have $A\mathcal{D}_k \subset \mathcal{D}_{k+1}$ and $r_{k+1} \perp \mathcal{D}_{k+1}$ it follows that r_{k+1} is A–orthogonal to $\mathcal{D}_k$. This allows us to eliminate most of the β_{kj} as follows: Using (14)

$$r_k^T r_{j+1} = r_k^T(r_j - \alpha_j Ad_j) = r_k^T r_j - \alpha_j r_k^T Ad_j$$

we have due to $r_k \perp \mathcal{D}_k$

$$r_k^T Ad_j = \begin{cases} \frac{1}{\alpha_k} r_k^T r_k & k = j \\ -\frac{1}{\alpha_{k-1}} r_k^T r_k & k = j+1 \\ 0 & \text{else.}\end{cases}$$

Inserting this into (11) we receive

$$\beta_{kj} = \begin{cases} \frac{1}{\alpha_{k-1}} \frac{r_k^T r_k}{d_{k-1}^T Ad_{k-1}} & k = j+1 \\ 0 & k > j+1.\end{cases}$$

For simplification we now write

$$\beta_k := \beta_{k,k-1}$$

and using (8) and (13) we receive

$$\beta_k = \frac{d_{k-1}^T Ad_{k-1}}{d_{k-1}^T r_{k-1}} \frac{r_k^T r_k}{d_{k-1}^T Ad_{k-1}} = \frac{r_k^T r_k}{d_{k-1}^T r_{k-1}} = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}.$$

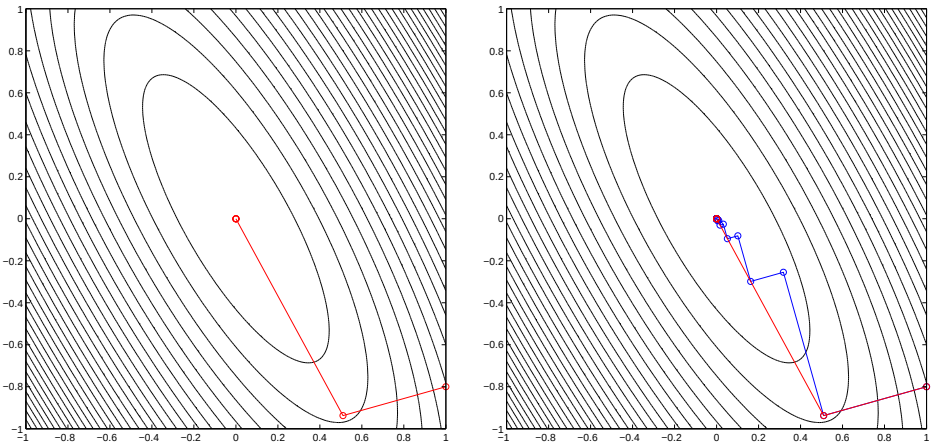
In total we have now:

Conjugate gradient. Let $f : \mathbb{R}^n \to \mathbb{R}$ with $f(x) = \frac{1}{2}x^T Ax - x^T b + c$, where $A \in \mathbb{R}^{n\times n}$ symmetric positive definite, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. $\min_x f(x)$ is solved in n steps by the following iterative scheme using a starting value $x_0 \in \mathbb{R}^n$:

$d_0 = r_0 = b - Ax_0$
iterate for $k = 0, \ldots, n-1$

$$\begin{aligned}\alpha_k &= \frac{r_k^T r_k}{d_k^T Ad_k}\\ x_{k+1} &= x_k + \alpha_k d_k\\ r_{k+1} &= r_k - \alpha_k Ad_k\\ \beta_{k+1} &= \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}\\ d_{k+1} &= r_{k+1} + \beta_{k+1} d_k\end{aligned}$$

The earlier example showing the zig–zagging with gradient descent is now nicely solved using CG (red path, blue path shows gradient descent):



More details and proofs can be found in Jonathan Shewchuk's excellent *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*, available online.

3.6 CG variants

Preconditioning

To improve convergence speed of the CG method, a technique called **Preconditioning** can be applied. It attempts to improve the condition of the matrix involved. Instead of solving $Ax = b$, one solves instead

$$M^{-1}Ax = M^{-1}b$$

using a matrix M, such that the condition of $M^{-1}A$ is improved over A. The CG method is easily modified to account for this, needing only one application of M^{-1} to the residual.

Several choices are available, from the non–sensical but perfect $M = A$ to the simple diagonal preconditioning ($M = \text{diag}(A)$) and many more advanced methods.

More details can be found in Jonathan Shewchuk's excellent *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*, available online.

CG on normal equation

The CG method derived in the previous section requires the matrix A to be square, symmetric and positive definite. If A does not have these properties, CG is still applicable via the normal equation. To be precise:

Let $A \in \mathbb{R}^{m \times n}$, with $m > n$ and A full rank, and $b \in \mathbb{R}^m$, then the least squares problem

$$\min_x \tfrac{1}{2}\|Ax - b\|^2$$

can be solved by applying the CG method to the normal equation $A^TAx = A^Tb$, as A^TA is symmetric and positive definite.

Convergence is unfortunately slower, but each iteration can still be implemented efficiently by never forming the matrix A^TA itself, but for example computing A^TAd in two steps, Ad and then A^TAd. Furthermore d^TA^TAd can be stably computed as $\langle Ad, Ad\rangle$.

Nonlinear CG

The previous section studied minimization of a quadratic form f, the corresponding CG variant is also often called "linear CG". The CG method can be adapted to general nonlinear functions f by changing the computation of the residual r_k, of the step size α_k and of the β_k. The nice convergence properties of linear CG are unfortunately mostly lost by these steps. Nonlinear CG is a popular method nevertheless (for lack of better alternatives).

Without going into further details, in nonlinear CG r_k is always computed as $r_k = -\nabla f(x_k)$, α_k needs to be computed via a line search algorithm like the Newton method, and there are several options to compute β_k, the easiest one is called the "Fletcher–Reeves formula" and is exactly the same as in linear CG.

3.7 Tikhonov regularization

Before introducing Tikhonov regularization, let us first revisit linear equation systems. Consider

$$Ax = b$$

for some matrix $A \in \mathbb{R}^{m\times n}$, unknowns $x \in \mathbb{R}^n$ and a right-hand side $b \in \mathbb{R}^m$. Usually, when we are modeling real problems, there is some unknown error ε, making our life hard:

$$Ax + \varepsilon = b.$$

In particular, if A is *ill-conditioned*, there can be issues, as outlined below.

Ill-conditioning. Let us assume for this paragraph that A is nicely square (i.e. $m = n$) and invertible. Then we know

$$A^{-1}b = A^{-1}(Ax + \varepsilon) = x + A^{-1}\varepsilon.$$

Using the induced matrix norm $\|A\|_2 := \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_{\max}(A)$, where $\sigma_{\max}(A)$ is the maximum singular value of A, we can estimate the size of the resulting error $A^{-1}\varepsilon$ by

$$\|A^{-1}\varepsilon\|_2 \leq \|A^{-1}\|_2 \, \|\varepsilon\|_2.$$

Hence, if $\|A^{-1}\|_2$ is big, the resulting error can be quite large, even if the original error ε is small. To elaborate on this: let $A = U\Sigma V^T$ be the SVD of A with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, then we have

$$A^{-1} = V\Sigma^{-1}U^T \qquad \text{with} \quad \Sigma^{-1} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n}\right).$$

If the condition of A, i.e. $\operatorname{cond}(A) := \frac{\sigma_1}{\sigma_n}$, is large (which is the case in almost all inverse problems), that is, if for example we have $\sigma_1 = 1$ and $\sigma_n <<<< 1$, then we have $\|A^{-1}\|_2 = \frac{1}{\sigma_n} >>>> 1$.

Coping with ill-conditioning. Let now A be square or non-square. Using the SVD $A = U\Sigma V^T$ with $U = (u_1, \ldots, u_n) \in \mathbb{R}^{m\times n}$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$ and $V = (v_1, \ldots, v_n) \in \mathbb{R}^{n\times n}$, we have

$$x = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i,$$

provided that $\sigma_i \neq 0$ for all $i = 1, \ldots, n$. If we stop summation at index $k < n$, we are computing the *truncated SVD solution*

$$\hat{x} = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i,$$

or using matrix notation:

$$\hat{x} = \mathcal{T}_k(b) := V \Sigma_k^+ U^T b$$

with $\Sigma_k^+ := \operatorname{diag}\big(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_k}, 0, \ldots, 0\big)$. We then have

$$\mathcal{T}_k(b) = V \Sigma_k^+ U^T (Ax + \varepsilon) = V \Sigma_k^+ \Sigma V^T x + V \Sigma_k^+ U^T \varepsilon,$$

where the first term of the sum approximates x and the second term is the resulting error. Using the induced matrix norm as above, we have

$$\big\|V \Sigma_k^+ U^T \varepsilon\big\|_2 \leq \big\|V \Sigma_k^+ U^T\big\|_2 \, \|\varepsilon\|_2 = \big\|\Sigma_k^+\|_2 \, \|\varepsilon\|_2 = \frac{1}{\sigma_k} \|\varepsilon\|_2.$$

Since $\sigma_1^{-1} \leq \ldots \leq \sigma_k^{-1}$, we get less and less noise amplification as k decreases.

Tikhonov regularization. Using $Ax + \varepsilon = b$ as before, the *Tikhonov regularized solution* is defined as

$$T_\alpha(b) := \arg\min_{z \in \mathbb{R}^n} \|Az - b\|_2^2 + \alpha \|z\|_2^2,$$

where $\alpha > 0$ is called *regularization parameter*.

$T_\alpha(b)$ thus balances a small residual $\big\|AT_\alpha(b) - b\big\|_2^2$ versus a small $\big\|T_\alpha(b)\big\|_2^2$, and $\alpha > 0$ "controls" the balance.

Using the SVD of $A = U \Sigma V^T$, we have

$$T_\alpha(b) = V D_\alpha^+ U^T b,$$

where

$$D_\alpha^+ := \operatorname{diag}\left(\frac{\sigma_1}{\sigma_1^2 + \alpha}, \ldots, \frac{\sigma_n}{\sigma_n^2 + \alpha}\right).$$

Proof: $V = (v_1, \ldots, v_n)$ is orthogonal, hence we can write $T_\alpha(b)$ as a linear combination of the v_i: $T_\alpha(b) = \sum_{i=1}^n a_i v_i$ with coefficients $(a_i)_{i=1}^n =: a$. In short notation this is $T_\alpha(b) = Va$. Now define $b' := U^T b$, then we have

$$\begin{aligned}
&\big\|AT_\alpha(b) - b\big\|_2^2 + \alpha \big\|T_\alpha(b)\big\|_2^2 \\
&\quad = \big\|U \Sigma V^T V a - U U^T b\big\|_2^2 + \alpha \|Va\|_2^2 \\
&\quad = \big\|\Sigma a - b'\big\|_2^2 + \alpha \|a\|_2^2 \\
&\quad = \sum_{i=1}^{\operatorname{rank}(A)} (\sigma_i a_i - b_i')^2 + \sum_{i=\operatorname{rank}(A)+1}^{n} (b_i')^2 + \alpha \sum_{i=1}^{n} a_i^2 \\
&\quad = \sum_{i=1}^{\operatorname{rank}(A)} (\sigma_i^2 + \alpha) \left(a_i^2 - 2 \frac{\sigma_i b_i'}{\sigma_i^2 + \alpha} a_i\right) + \alpha \sum_{i=\operatorname{rank}(A)+1}^{n} a_i^2 + \sum_{i=1}^{n} (b_i')^2 \\
&\quad = \sum_{i=1}^{\operatorname{rank}(A)} (\sigma_i^2 + \alpha) \left(a_i - \frac{\sigma_i b_i'}{\sigma_i^2 + \alpha}\right)^2 + \alpha \sum_{i=\operatorname{rank}(A)+1}^{n} a_i^2 - \sum_{i=1}^{\operatorname{rank}(A)} \frac{(\sigma_i b_i')^2}{\sigma_i^2 + \alpha} + \sum_{i=1}^{n} (b_i')^2
\end{aligned}$$

Our task is now to choose a_i such that this term is minimized. The correct choice is obviously

$$a_i = \begin{cases} \frac{\sigma_i}{\sigma_i^2+\alpha}\, b'_i & \text{for } 1 \le i \le \operatorname{rank}(A) \\ 0 & \text{for } i > \operatorname{rank}(A) \end{cases}$$

This implies $a = D_\alpha^+ b'$, which concludes the proof. □

With Tikhonov regularization we thus have

$$T_\alpha(b) = V D_\alpha^+ U^T b = \sum_{i=1}^{\operatorname{rank}(A)} \left(\frac{\sigma_i}{\sigma_i^2 + \alpha} \right) (u_i^T b) v_i.$$

With truncated SVD we have

$$\mathcal{T}_k(b) = \sum_{i=1}^{k} \frac{1}{\sigma_i} (u_i^T b) v_i.$$

If we choose k such that $k = \max\{i : \sigma_i > \alpha\}$, then for small α these are essentially the same. Increasing α means placing less weight on the small singular values (which is equivalent to a Wiener filter).

Stacked form and normal equation. Another equivalent formulation is to interpret the Tikhonov minimization problem as two parts, $Ax = b$ and $x = 0$, for which we want to compute the least squares solution simultaneously. We can write this as a stacked equation

$$\underbrace{\begin{bmatrix} A \\ \sqrt{\alpha} I \end{bmatrix}}_{=:\tilde{A}} x = \underbrace{\begin{bmatrix} b \\ 0 \end{bmatrix}}_{=:\tilde{b}}, \qquad \text{or in short} \quad \tilde{A}x = \tilde{b}.$$

The least squares solution of this satisfies the normal equation

$$\tilde{A}^T \tilde{A} x = \tilde{A}^T \tilde{b},$$

where

$$\tilde{A}^T \tilde{A} = \begin{bmatrix} A^T & \sqrt{\alpha} I \end{bmatrix} \begin{bmatrix} A \\ \sqrt{\alpha} I \end{bmatrix} = A^T A + \alpha I,$$

and

$$\tilde{A}^T \tilde{b} = \begin{bmatrix} A^T & \sqrt{\alpha} I \end{bmatrix} \begin{bmatrix} b \\ 0 \end{bmatrix} = A^T b.$$

Hence this is equivalent to

$$(A^T A + \alpha I) x = A^T b.$$

We can use this normal equation to compute the Tikhonov regularized solution conveniently as

$$T_\alpha(b) = \left(A^T A + \alpha I \right)^{-1} A^T b.$$

Generalizations. Very similar formulas can be developed for the following problems:

$$\arg\min_{z \in \mathbb{R}^n} \|Az - b\|_2^2 + \alpha \|z - x^*\|_2^2,$$

where we want the solution to be close to some known x^*. Or we want to enforce smoothness via

$$\arg\min_{z\in\mathbb{R}^n} \|Az-b\|_2^2 + \alpha\|Lz\|_2^2,$$

where L is a finite differences operator (e.g. forward differences). Finally,

$$\arg\min_{z\in\mathbb{R}^n} \|Az-b\|_2^2 + \alpha\big\|L(z-x^*)\big\|_2^2.$$

Choosing the regularization parameter α. There is, unfortunately, no general solution on how to choose α.

However, there are some imperfect approaches. One of them is the *L-curve method*: choose candidates $0 < \alpha_1 < \ldots < \alpha_M < \infty$, compute $T_{\alpha_m}(b)$ for $1 \le m \le M$ and plot the 2D curve $\big(\log\big\|AT_{\alpha_m}(b)-b\big\|,\ \log\big\|T_{\alpha_m}(b)\big\|\big)$ for all $1 \le m \le M$. This curve typically has an L-shape, so we choose a point in the "corner" of the L. While this is a handy method for choosing α, it has high computational effort (having to compute all the $T_\alpha(b)$), and the chosen α might not yield a solution of the desired properties...

3.8 Newton method

We investigate the minimization problem for $f : \mathbb{R}^n \to \mathbb{R}$, where:

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{with} \quad f(x) = \tfrac{1}{2}\|g(x)\|_2^2,$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$ (see for example the Case Study from section 2.7).

If $g(x) = Ax - b$ with $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, this is a linear least squares problem. There exist a lot of solution approaches from linear algebra (for example normal equation, singular value decomposition, ART and others).

Problem: If g is non–linear, we need other methods.

Assumption: g is twice continuously differentiable and f fulfills a sufficient condition for a local minimum (see section 3.3). We thus want to find $x^* \in \mathbb{R}^n$ such that $\nabla f(x^*) = 0$ and $H_f(x^*)$ positive definite ($H_f(x^*)$ is the Hessian of f at x^*). Computing $\nabla f(x)$, we get

$$F(x) := \nabla f(x) = J_g(x)^T g(x) \in \mathbb{R}^n,$$

where $J_g(x) \in \mathbb{R}^{m\times n}$ is the Jacobi matrix of $g : \mathbb{R}^n \to \mathbb{R}^m$. We have to determine $x^* \in \mathbb{R}^n$ with

$$F(x^*) = J_g(x^*)^T g(x^*) = 0.$$

This is a system of n non–linear equations. One general method to find a solution is:

Newton's method. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ continuously differentiable. To find a solution $x^* \in \mathbb{R}^n$ to

$$F(x) = 0$$

we

1. Choose an initial value $x_0 \in \mathbb{R}^n$ and approximate F with its first-order Taylor expansion in x_0 (in order to have a linear problem in x)

$$F(x) \approx F(x_0) + J_F(x_0)(x - x_0).$$

2. Find a solution x_1 of (the linear problem)

$$F(x_0) + J_F(x_0)(x - x_0) = 0.$$

3. Having found x_1, determine a solution x_2 of

$$F(x_1) + J_F(x_1)(x - x_1) = 0.$$

4. Iterating this procedure we obtain a sequence of vectors $x_k \in \mathbb{R}^n$, where

$$x_{k+1} = x_k + \Delta x_k \qquad \text{with} \quad \Delta x_k = -J_F(x_k)^{-1} F(x_k).$$

Of course we have to assume that the Jacobi matrices $J_F(x_k)$ are all invertible. In general x_k then converges to x^*.

We now investigate Newton's method for the one-dimensional case to give a graphical view of the iterative procedure.

Theorem. *Let $F : [a, b] \to \mathbb{R}$ twice differentiable and convex with $F(a) < 0$, $F(b) > 0$. Then:*

1. *There exists exactly one $x^* \in (a, b)$ such that $F(x^*) = 0$.*
2. *If $x_0 \in [a, b]$ such that $F(x_0) > 0$, then*

$$x_{k+1} := x_k - \frac{F(x_k)}{F'(x_k)}$$

is well-defined (in particular $F'(x_k) > 0$) and the sequence $(x_k)_{k \in \mathbb{N}_0}$ converges to x^* monotonically decreasing.

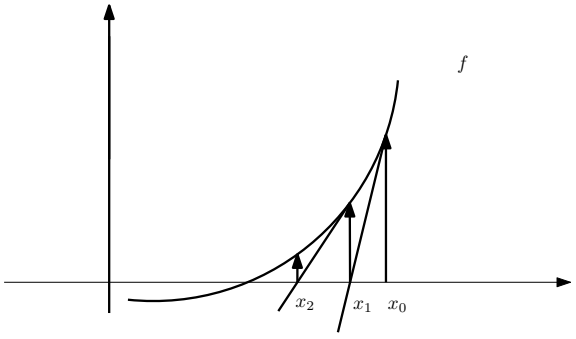
3. *If $F'(x^*) \geq C > 0$ and $F''(x) \leq K$ for all $x \in (x^*, b)$, then*

$$|x_{k+1} - x_k| \leq |x^* - x_k| \leq \frac{K}{2C} |x_k - x_{k-1}|^2.$$

Remarks:

- The convergence statement of the theorem is true also if $F(a) > 0, F(b) < 0$ or if F is concave.
- The third statement in the theorem says that Newton's method is of quadratic convergence. (That means if x_k and x_{k-1} have l identical decimals, then x_{k+1} and x_k have $2l$ identical decimals, provided $\frac{K}{2C} \approx 1$.)

In the one–dimensional case, Newton's method can be visualized by iteratively approximating F with its tangent at x_k and seeking the zero of it:



For the n–dimensional case, we can summarize:

Theorem (Newton's method)**.** *Let $\Omega \subset \mathbb{R}^n$ open, $F : \Omega \to \mathbb{R}^n$ continuously differentiable and Lipschitz-continuous. Assume there exists $x^* \in \Omega$ with $F(x^*) = 0$ and that $J_F(x^*)^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ exists.*

*The **Newton iteration** (where possible) is defined as*

$$x_0 \in \Omega, \qquad x_{n+1} := x_n - J_F(x_n)^{-1} F(x_n).$$

Then there exists $\delta > 0$ such that $(x_n)_{n\in\mathbb{N}_0}$ exists and converges to x^* for every $\|x_0 - x^*\| < \delta$.

Now we apply Newton's method to our previous minimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \text{with} \quad f(x) = \tfrac{1}{2}\|g(x)\|_2^2,$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$. We have to compute $J_F(x)$, where $F(x) = \nabla f(x) = J_g(x)^T g(x)$. With the product rule we get

$$J_F(x) = g''(x)^T g(x) + J_g(x)^T J_g(x),$$

where $g''(x)$ is the derivative of $J_g : \mathbb{R}^n \to \mathbb{R}^{m\times n}$, that is $g'' : \mathbb{R}^n \to (\mathbb{R}^n \times \mathbb{R}^{m\times n})$.

We see that $g''(x)$ is not easy to deal with. Thus it is often assumed that $g(x) \approx 0$, i.e. x is near to the solution of our problem. This leads to the following method:

Gauss-Newton method. Let $g : \mathbb{R}^n \to \mathbb{R}^m$ be twice continuously differentiable and assume that $J_g(x)^T J_g(x) > 0$ for all $x \in \mathbb{R}^n$ (we have $J_g(x)^T J_g(x) \in \mathbb{R}^{n\times n}$). Then the solution $x^* \in \mathbb{R}^n$ to

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|g(x)\|_2^2$$

can be computed iteratively for $k = 0, 1, 2, \ldots$ and an initial value $x_0 \in \mathbb{R}^n$ by

$$x_{k+1} = x_k + \Delta x_k \qquad \text{with} \quad \Delta x_k = -\big(J_g(x_k)^T J_g(x_k)\big)^{-1} J_g(x_k)^T g(x_k).$$

(recall that $F(x) = J_g(x)^T g(x)$.)

In the context of least squares problems we can interpret Gauss-Newton differently, see the normal equation in section 1.11: $\Delta x_k = x_{k+1} - x_k$ is the least squares solution of the least squares problem

$$\min_{\Delta x_k} \big\|J_g(x_k)\Delta x_k + g(x_k)\big\|_2^2.$$

That means the original non–linear least squares problem has been replaced by a series of linear least squares problems.

3.9 Case Study (Part II)

The tracking problem from section 2.7 (Case Study, Part I) was formulated as

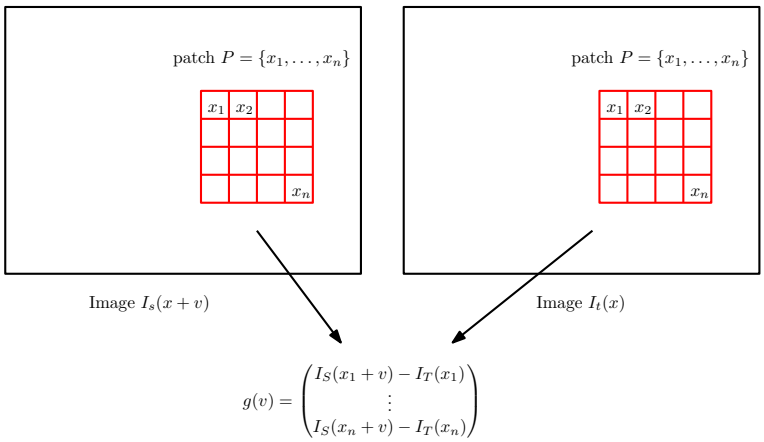
$$\min_{v\in\mathbb{R}^2} \int_P \big|I_s(x+v) - I_t(x)\big|^2,$$

where I_s, I_t denote the source and target images as functions $\Omega \subset \mathbb{R}^2 \to \mathbb{R}$, and $P \subset \Omega$ a patch wherein v was assumed constant. Discretizing P into n pixels, $P = \{x_1, \dots, x_n\}$, the problem reduced to a linear least squares problem via the Taylor expansion,

$$\min_{v \in \mathbb{R}^2} \|g(v)\|_2^2 \qquad \text{with} \quad g(v) = \begin{pmatrix} I_s(x_1) + \nabla I_s(x_1)^T v - I_t(x_1) \\ \vdots \\ I_s(x_n) + \nabla I_s(x_n)^T v - I_t(x_n) \end{pmatrix}.$$

Instead we now directly treat the non–linear problem

$$\min_{v \in \mathbb{R}^2} \|g(v)\|_2^2 \qquad \text{with} \quad g(v) = \begin{pmatrix} I_s(x_1 + v) - I_t(x_1) \\ \vdots \\ I_s(x_n + v) - I_t(x_n) \end{pmatrix}.$$



Using Gauss–Newton, an iteration step for v reads as

$$v_{k+1} = v_k + \Delta v_k \qquad \text{with} \quad \Delta v_k = -\big(J_g(v_k)^T J_g(v_k)\big)^{-1} J_g(v_k) g(v_k),$$

where

$$J_g(v) = \begin{pmatrix} \nabla I_s(x_1 + v)^T \\ \vdots \\ \nabla I_s(x_n + v)^T \end{pmatrix} = \begin{pmatrix} \partial_1 I_s(x_1 + v) & \partial_2 I_s(x_1 + v) \\ \vdots & \vdots \\ \partial_1 I_s(x_n + v) & \partial_2 I_s(x_n + v) \end{pmatrix}.$$

Put differently, here each iteration step for Δv_k is the least squares solution of

$$\min_{\Delta v_k} \big\| J_g(v_k) \Delta v_k - g(v_k) \big\|_2^2,$$

while for the linear approximation in section 2.7 (Part I) we solved just one linear least squares problem,

$$\min_{v} \big\| J_g(0) v - g(0) \big\|_2^2.$$

3.10 Fixed point iteration

Definition (Fixed point)**.** *Let $N \subset M \neq \emptyset$ be sets, $T : M \to N$. A solution $x^* \in N$ of the equation*

$$x = T(x)$$

*is called **fixed point** of T.*

Examples:

- $T_1 : \mathbb{R} \to \mathbb{R}$, $T_1(x) = 1 + x \quad \Rightarrow$ no fixed point.
- $T_2 : \mathbb{R} \to \mathbb{R}$, $T_2(x) = ax$ with $a \in \mathbb{R}$, $a \neq 1 \quad \Rightarrow$ exactly one fixed point.
- $T_3 : \mathbb{R} \to \mathbb{R}$, $T_3(x) = x \quad \Rightarrow$ infinitely many fixed points.

Definition (Fixed point iteration)**.** *Let (M, d) be a metric space, $T : M \to M$. The iteration scheme*

$$x_0 \in M, \qquad x_{n+1} = T(x_n)$$

*is called **fixed point iteration** of T.*

Examples: (continued)

- T_1 fixed point iteration does not converge.
- T_2 fixed point iteration does not converge for $|a| > 1$, it does converge for $|a| < 1$.
- T_3 fixed point iteration converges.

Definition (Lipschitz–continuous / contraction)**.** *Let (M, d) be a metric space, $A \subset M$ and $T : A \to M$. T is called **Lipschitz–continuous** if there exists $L > 0$ such that*

$$d\big(T(x), T(y)\big) \leq L \cdot d(x, y) \qquad \forall x, y \in A.$$

*If $L \in (0, 1)$, then T is called a (proper) **contraction**.*

Theorem (Banach fixed point theorem)**.** *Let (X, d) be a complete metric space, $M \subset X$, $M \neq \emptyset$ and $T : M \to M$. If M is closed in X and T a (proper) contraction, then there exists exactly one fixed point $x^* \in M$ and the fixed point iteration converges.*

Furthermore, if $L \in (0, 1)$ is a Lipschitz constant of T, then

$$d(x_k, x^*) \leq \frac{L}{1 - L} d(x_{k-1}, x_k).$$

Proof: Let $x_0 \in M$, $L \in (0, 1)$ a Lipschitz constant of T. We have for the fixed point iteration $x_{n+1} = T(x_n)$

$$d(x_k, x_{k+1}) = d\big(T(x_{k-1}), T(x_k)\big) \leq L d(x_{k-1}, x_k) \qquad \forall k \in \mathbb{N},$$

and thus

$$d(x_k, x_{k+1}) \leq L^k d(x_0, x_1) \qquad \forall k \in \mathbb{N}.$$

Hence for all $k, m \in \mathbb{N}$

$$\begin{aligned} d(x_k, x_{k+m}) &\leq d(x_k, x_{k+1}) + \ldots + d(x_{k+m-1}, x_{k+m}) \\ &\leq L^k d(x_0, x_1) + \ldots + L^{k+m-1} d(x_0, x_1) = L^k \left(\sum_{j=0}^{m-1} L^j \right) d(x_0, x_1) \\ &\leq \frac{L^k}{1 - L} \, d(x_0, x_1) \end{aligned}$$

as $L < 1$. That means (x_k) is a Cauchy sequence in M. Since X is complete and M closed in X, $\lim_{k\to\infty} x_k = x^*$ for some $x^* \in M$. T is a contraction and thus continuous, implying

$$x^* = \lim_{k\to\infty} x_k = \lim_{k\to\infty} T(x_{k-1}) = T(x^*).$$

If $y^* \in M$ is another fixed point of T, we have

$$d(x^*, y^*) = d\big(T(x^*), T(y^*)\big) \le L d(x^*, y^*),$$

that is $(1-L)d(x^*, y^*) \le 0 \;\Rightarrow\; d(x^*, y^*) = 0 \;\Rightarrow\; x^* = y^*$.

To show the convergence rate, we have

$$d(x_k, x_{k+m}) \le L \left(\sum_{j=0}^{m-1} L^j \right) d(x_{k-1}, x_k) \le \frac{L}{1-L} d(x_{k-1}, x_k)$$

and thus

$$d(x_k, x^*) = \lim_{m\to\infty} d(x_k, x_{k+m}) \le \frac{L}{1-L} d(x_{k-1}, x_k).$$

□

Another fixed point theorem with less restrictions, but not easy to prove, is:

Theorem (Brouwer fixed point theorem)**.** *Let $M \subset \mathbb{R}^n$ be convex and compact in $\mathbb{R}^n$, $M \neq \emptyset$, and $T : M \to M$ continuous. Then there exists at least one fixed point $x^* \in M$ of T.*

Application: Let $F : \mathbb{R}^n \to \mathbb{R}^n$, we search for $x^* \in \mathbb{R}^n$ with $F(x^*) = 0$.

Idea: fixed point iteration with $T : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$F(x^*) = 0 \quad \Leftrightarrow \quad T(x^*) = x^*.$$

Examples:

- $T(x) := x - F(x)$
- $T(x) := x - \alpha F(x)$ (linear relaxation)
- $T(x) := x - J_F(x)^{-1} F(x)$ (Newton method)

Theorem (Newton method)**.** *Let $\Omega \subset \mathbb{R}^n$ open, $F : \Omega \to \mathbb{R}^n$ continuously differentiable and Lipschitz-continuous. Assume there exists $x^* \in \Omega$ with $F(x^*) = 0$ and that $J_F(x^*)^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ exists. Define now the Newton iteration (where possible)*

$$x_0 \in \Omega, \qquad x_{n+1} := x_n - J_F(x_n)^{-1} F(x_n).$$

Then there exists $\delta > 0$ such that $(x_n)_{n\in\mathbb{N}_0}$ exists and converges to x^* for every $\|x_0 - x^*\| < \delta$.

4 Probability theory

4.1 Basics of probability

Example: Coin toss experiment. What are the possible outcomes of this experiment? You can throw head (H) or tails (T) - i.e. the **sample space** is $\Omega_c = \{H, T\}$. If you are interested how a particular experiment turned out, you want to know the **event** that happened, e.g. whether the event $A = \{H\}$ happened.

Example: Throwing dice experiment. Sample space $\Omega_d = \{1, 2, 3, 4, 5, 6\}$. Potential event you are interested in $A = \{5, 6\}$ or $B = \{2, 4, 6\}$.

Definition (Sample space, event space)**.** *The set of all potential outcomes Ω of a random experiment is called **sample space**.*

*A subset $A \subset \Omega$ of the sample space is called **event**. The event A is said to happen if during the experiment an outcome $\omega \in A$ occurs. $\emptyset$ is the impossible event, Ω the certain event.*

*A set of events $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called **event space** or **σ-field** if it fulfils*

- $\Omega \in \mathcal{F}$,
- $A \in \mathcal{F} \;\Rightarrow\; A^c \in \mathcal{F}$,
- $A_i \in \mathcal{F}$ *for* $i \in \mathbb{N}$ *implies* $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$.

Here $\mathcal{P}(\Omega)$ denotes the power set of Ω (the set of all subsets of Ω).

Example: The event space of the coin and dice experiments are just

$$\mathcal{P}(\Omega_c) = \big\{\emptyset, \Omega_c, \{H\}, \{T\}\big\}$$

and $\mathcal{P}(\Omega_d)$. The event space being the power set of the sample space is typical for finite sample spaces.

Property: For every set of sets $\mathcal{A} \subset \mathcal{P}(\Omega)$ there exists a smallest σ-field $\sigma(\mathcal{A})$ that contains $\mathcal{A}$. We say that $\mathcal{A}$ **creates** $\sigma(\mathcal{A})$.

Definition (Borel-σ-field)**.** *Let (Ω, d) be a metric space. The σ–field created by the open sets of Ω is called the **Borel-σ-field** $\mathcal{B}(\Omega)$ of Ω.*

Properties:

- $\mathcal{B}(\mathbb{R})$ contains all intervals (a, b), $(a, b]$, $[a, b]$, $[a, b)$ where $a, b \in \mathbb{R}$, $a < b$, as well as $(-\infty, a]$, $a \in \mathbb{R}$.
- $\mathcal{B}(\mathbb{R}) \neq \mathcal{P}(\mathbb{R})$, but finding $A \subset \mathbb{R}$ with $A \notin \mathcal{B}(\mathbb{R})$ is hard.

Definition (Probability measure, probability space)**.** *Let Ω be the sample space, $\mathcal{F}$ the event space. A function $P : \mathcal{F} \to \mathbb{R}$ is called a **probability measure** if it fulfills*

- $P(A) \geq 0$ *for all* $A \in \mathcal{F}$,
- $P(\Omega) = 1$,

- *If $A_i \in \mathcal{F}$, $i \in \mathbb{N}$, are disjoint events, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$P\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} P(A_i).$$

*The triple $(\Omega, \mathcal{F}, P)$ is called **probability space**.*

Example:

- Coin toss: $(\Omega_c, \mathcal{P}(\Omega_c), P_c)$ is a probability space with $P_c : \mathcal{P}(\Omega_c) \to \mathbb{R}$, $P_c(\emptyset) = 0$, $P_c(\Omega_c) = 1$ and $P_c(\{H\}) = P_c(\{T\}) = \frac{1}{2}$.
- Dice: $(\Omega_d, \mathcal{P}(\Omega_d), P_d)$ is a probability space with $P_d : \mathcal{P}(\Omega_d) \to \mathbb{R}$, $P_d(\emptyset) = 0$, $P_d(\Omega_d) = 1$ and $P(\{i\}) = \frac{1}{6}$ for $i = 1, \ldots, 6$ and applying the third rule for all missing sets of $\mathcal{P}(\Omega_d)$.

Properties: Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then for $A, B \in \mathcal{F}$ we have

$$\begin{aligned} P(A) \in [0,1], \quad P(\emptyset) = 0, \\ P(A^c) = 1 - P(A), \\ P(A \setminus B) = P(A) - P(A \cap B), \\ P(A \cup B) = P(A) + P(B) - P(A \cap B), \\ P(A) \leq P(B) \quad \text{if } A \subset B. \end{aligned}$$

Definition (Counting rule)**.** *Let Ω be a finite set and $P : \mathcal{P}(\Omega) \to \mathbb{R}$ with $P(\{\omega\}) = \frac{1}{|\Omega|}$ for all $\omega \in \Omega$. Extending this canonically to a probability space $(\Omega, \mathcal{P}(\Omega), P)$ we get the **counting rule** for $A \in \mathcal{P}(\Omega)$*

$$P(A) = \frac{|A|}{|\Omega|}.$$

This is what we implicitly did with the dice experiment example.

Definition (Conditional probability)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. For $A, B \in \mathcal{F}$ with $P(B) \neq 0$, the term*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

*is called **conditional probability** of A under the condition B.*

Theorem (Total probability)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $(B_i)_{i\in\mathbb{N}} \subset \mathcal{F}$ be a sequence of pairwise disjoint events in $\mathcal{F}$ (i.e. $B_i \cap B_j = \emptyset$ for $i \neq j$), and $A \in \mathcal{F}$ with $A \subset \bigcup_{i\in\mathbb{N}} B_i$. Then*

$$P(A) = \sum_{i\in\mathbb{N}} P(B_i)\ P(A|B_i).$$

Theorem (Bayes)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $B_1, \ldots, B_n \in \mathcal{F}$ pairwise disjoint events and $A \in \mathcal{F}$ with $A \subset \bigcup_{i=1}^n B_i$. Then*

$$P(B_j|A) = \frac{P(B_j)\ P(A|B_j)}{\sum_{i=1}^n P(B_i)\ P(A|B_i)} \qquad \text{for all } j = 1, \ldots, n.$$

Definition (Independence)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two events $A, B \in \mathcal{F}$ are called* ***independent*** *if*

$$P(A \cap B) = P(A)P(B).$$

A family of events $A_i \in \mathcal{F}$, $i \in I$, is called ***stochastically independent****, if for every finite set $J \subset I$*

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j).$$

A family of events $A_i \in \mathcal{F}$, $i \in I$, is called ***pairwise independent*** *if*

$$P(A_i \cap A_j) = P(A_i)P(A_j) \qquad \text{for all } i \neq j.$$

4.2 Random variables

Definition (Measureable function)**.** *Let Ω, Ω' be two sample spaces with corresponding event spaces $\mathcal{F}, \mathcal{F}'$. A function $X : \Omega \to \Omega'$ which fulfills*

$$X^{-1}(A') \in \mathcal{F} \qquad \forall A' \in \mathcal{F}'$$

is called $\mathcal{F}$-$\mathcal{F}'$ ***measurable*** *or just measurable.*

Measureable functions play "nice" with event spaces, just like linear functions play "nice" with linear spaces in linear algebra.

Definition (Random variable, distribution)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is called a* ***random variable****, if it is $\mathcal{F}$-$\mathcal{B}(\mathbb{R})$ measurable.*

The probability measure $P_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ with $P_X := P \circ X^{-1}$ is called ***distribution*** *of X.*

For $B \in \mathcal{B}(\mathbb{R})$ this means:

$$P_X(B) = P\big(X^{-1}(B)\big) = P\big(\{\omega \in \Omega :\ X(\omega) \in B\}\big).$$

Property: For finite or countably infinite Ω and $\mathcal{F} = \mathcal{P}(\Omega)$, every mapping $X : \Omega \to \mathbb{R}$ is a random variable.

Example 1: In the coin toss experiment, the mapping $X : \Omega_c \to \mathbb{R}$ with

$$X(\omega) = \begin{cases} c & \text{for } \omega \in \{H\} \\ d & \text{else} \end{cases}$$

with $c, d \in \mathbb{R}$ is a random variable.

This type of experiment (only two distinct events) is called **Bernoulli experiment**. Or more formally: $\Omega = A \cup A^c$ for some set A, $\mathcal{F} = \mathcal{P}(\Omega)$ and $X : \Omega \to \{0, 1\}$, where $X(A) = 1$ and $X(A^c) = 0$. We now have for some $p \in (0, 1)$

$$P(A) = P\big(\{\omega \in \Omega :\ X(\omega) = 1\}\big) = P(X = 1) = p,$$
$$P(A^c) = P\big(\{\omega \in \Omega :\ X(\omega) = 0\}\big) = P(X = 0) = 1 - p.$$

Definition (Discrete random variable)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable with at most countably infinite many values, $X : \Omega \to \{x_i\}$, $i \in I$ where $|I| \leq |\mathbb{N}|$, is called a* ***discrete random variable****. The distribution P_X is already uniquely defined by the values $P(X = x_i)$, $i \in I$. The function*

$$f_X : \{x_i\} \to \mathbb{R}, \qquad f_X(x_i) = P(X = x_i)$$

is called ***probability mass function*** *(in short:* ***pmf****) of X.*

A Bernoulli experiment is thus a discrete random variable.

Example 2: If we perform a Bernoulli experiment (as defined above) n times, we can model this using a discrete random variable $X : \Omega \to \{0, 1, 2, \ldots, n\}$, where X tells you the number of times the event A happened. The distribution P_X is then defined by

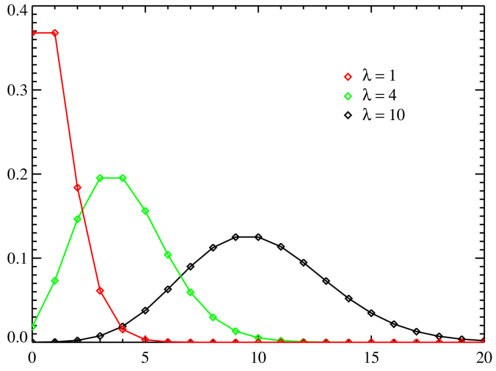
$$P(X = i) = \binom{n}{i} p^i \ (1-p)^{n-i},$$

and is called the **Binomial distribution** $B(n, p)$.

Example 3: A distribution relevant for emission tomography (ET) is the **Poisson distribution** $Poi(\lambda)$. For some $\Omega, \mathcal{F}$, the random variable $X : \Omega \to \mathbb{N}_0$ is supposed to count "rare" events (like the radioactive decay). The distribution P_X is then defined via

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!} \qquad \text{for } i \in \mathbb{N}_0,$$

where $\lambda > 0$ specifies the expected number of events. The probability mass function for several values of λ can be visualized as follows:



Another way to define a distribution is the to use the cumulative distribution function.

Definition (Cumulative distribution function)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : \Omega \to \mathbb{R}$ a random variable. The function $F_X : \mathbb{R} \to \mathbb{R}$,*

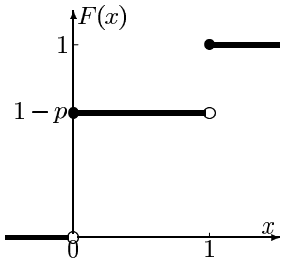
$$F_X(x) := P(X \leq x) = P\big(\{\omega \in \Omega : \ X(\omega) \leq x\}\big)$$

is called ***(cumulative) distribution function*** *(in short:* ***cdf****) of X.*

Comments:

- F_X is well defined, as $\{\omega \in \Omega : \ X(\omega) \leq x\} \in \mathcal{F}$ due X being measurable.
- F_X uniquely defines the distribution P_X of X, as $\mathcal{B}(\mathbb{R})$ is created by $(-\infty, a]$, $a \in \mathbb{R}$.
- F_X is monotonously non-decreasing ($x \leq y \ \Rightarrow \ F_X(x) \leq F_X(y)$), $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to\infty} F_X(x) = 1$ and F_X is right-continuous.

The cumulative distribution function for our earlier Bernoulli experiment is:



If our random variable is not discrete, we need a slightly modified concept instead of a probability mass function:

Definition (Probability density function)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : \Omega \to \mathbb{R}$ a random variable. A function $f_X : \mathbb{R} \to \mathbb{R}$ is called* ***(probability) density function*** *(in short:* ***pdf****), if it fulfills*

- $f_X(x) \geq 0$ *for all* $x \in \mathbb{R}$.
- f_X *integrable and* $\int_{\mathbb{R}} f_X(x)dx = 1$.
- $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ *for all* $x \in \mathbb{R}$.

X is then also called ***continuous random variable****.*

Properties: Let X be a continuous random variable with cdf F_X and pdf f_X.

- If F_X is continuously differentiable almost everywhere, then

$$f_X(x) = F_X'(x) \qquad \text{for almost all } x \in \mathbb{R}.$$

- We have for $a, b \in \mathbb{R}$, $a < b$

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(t)dt = P(a \leq X \leq b)$$

as well as

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(t)dt = 0.$$

Definition (Identically distributed random variables)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two random variables X, Y are called **identically distributed** (in short $X \sim Y$), if the distributions P_X, P_Y are equal, i.e.*

$$P\big(X^{-1}(B)\big) = P\big(Y^{-1}(B)\big) \qquad \text{for all } B \in \mathcal{B}(\mathbb{R}),$$

or equivalently

$$F_X(z) = F_Y(z) \qquad \text{for all } z \in \mathbb{R}.$$

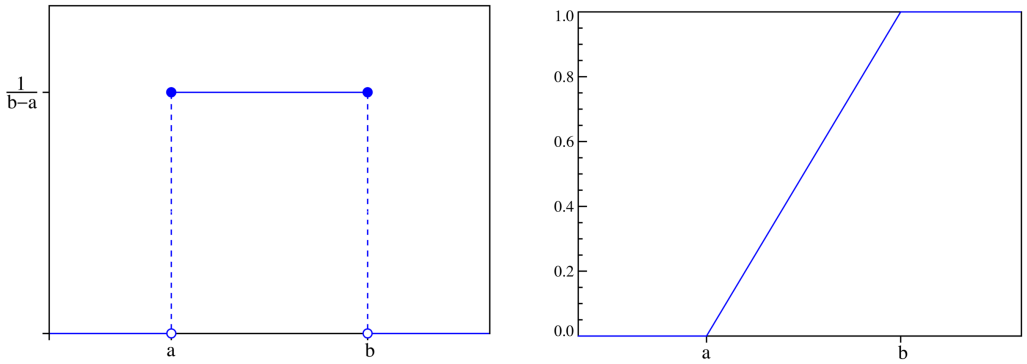
Example 4: Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable X is called **uniformly distributed** in (a, b) for $a, b \in \mathbb{R}$, $a < b$, in short $X \sim U(a, b)$ if it has the density function

$$f_X(x) = \frac{1}{b-a}\chi_{(a,b)}(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \text{else.} \end{cases}$$

The distribution function is

$$F_X(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & x \in (a, b) \\ 1 & x \ge b. \end{cases}$$

Here are graphs of the pdf and cdf of the uniform distribution:

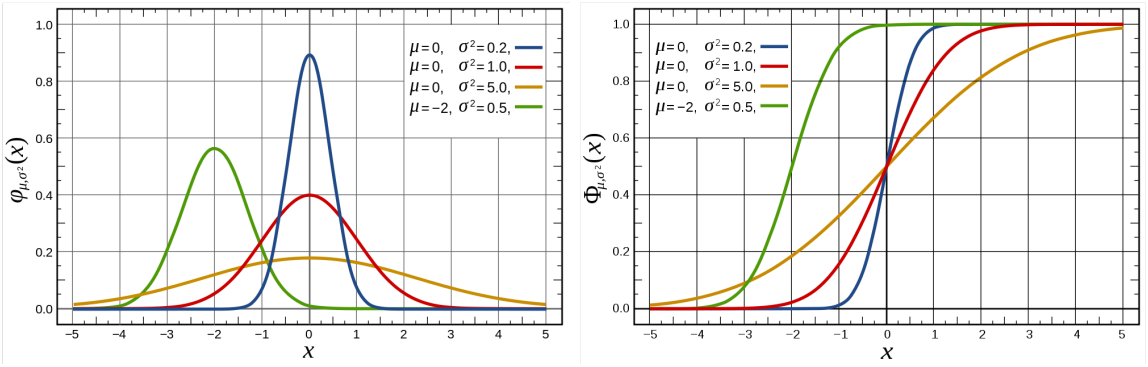


Example 5: Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable X is called **normally distributed** or **Gaussian**, in short $X \sim \mathcal{N}(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$, $\sigma^2 > 0$, if it has the density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}.$$

The special case $\mathcal{N}(0, 1)$ is called the **standard normal distribution**. There is no closed formula for the distribution function, but it can be computed.

Plots of the probability density function and cumulative distribution function of $\mathcal{N}(\mu, \sigma^2)$ for several values of μ, σ^2 are shown below:



Definition (Random vector)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be random variables. The mapping $(X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ is $\mathcal{F}$–$\mathcal{B}(\mathbb{R}^n)$ measurable and is called* ***random vector****. $P_{X_1,\ldots,X_n} : \mathcal{B}(\mathbb{R}^n) \to \mathbb{R}$, $P_{X_1,\ldots,X_n} = P \circ (X_1, \ldots, X_n)^{-1}$ is then called* ***joint distribution*** *of $X_1, \ldots, X_n$.*

*Alternatively in short–hand notation, $\mathbf{X} = (X_1, \ldots, X_n)^T$, $\mathbf{X} : \Omega \to \mathbb{R}^n$ and $P_{\mathbf{X}} = P \circ \mathbf{X}^{-1}$. $\mathbf{X}$ is then also called n–****dimensional random variable****.*

The cumulative distribution function for such a random variable is then $F_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}$ with

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &:= P(\mathbf{X} \leq \mathbf{x}) \\ &:= P\big(\{\omega \in \Omega : \ X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}\big) \end{aligned}$$

for $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

Example 6: Tossing a coin two times. $\Omega = \{H, T\} \times \{H, T\}$. If we denote X as the number of heads and Y the number of tails, then our 2–dimensional random variable is

$$(X, Y)^T : \Omega \to \{0, 1, 2\} \times \{0, 1, 2\}.$$

If $P(H) = p$ for $p \in (0, 1)$, then we have

$$P(X = 0, Y = 2) = (1 - p)^2, \quad P(X = 1, Y = 1) = 2p(1 - p), \quad P(X = 2, Y = 0) = p^2,$$

and all other $P(X = i, Y = j) = 0$, which defines us the distribution of $(X, Y)^T$.

Definition (n–dimensional continuous distribution)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ is called n–****dimensionally continuously distributed****, if there exists a non–negative, integrable function $f_{X_1,\ldots,X_n} : \mathbb{R}^n \to \mathbb{R}$ with*

$$F_{X_1,\ldots,X_n}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n$$

*for $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. $f_{X_1,\ldots,X_n}$ is called n–****dimensional density function****. At $\mathbf{x} \in \mathbb{R}^n$ where $f_{X_1,\ldots,X_n}$ is continuous, we have*

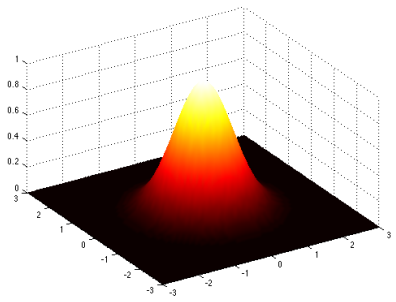
$$f_{X_1,\ldots,X_n}(\mathbf{x}) = \frac{\partial^n F_{X_1,\ldots,X_n}(\mathbf{x})}{\partial x_1 \cdots \partial x_n}.$$

Example 7: n–dimensional Gaussian distribution $\mathcal{N}(\mu, \mathbf{C})$. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathbf{X} = (X_1, \ldots, X_n)^T$ a random vector. Let $\mu \in \mathbb{R}^n$ and $\mathbf{C} \in \mathbb{R}^{n \times n}$ a symmetric positive definite matrix. We have $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{C})$, if $\mathbf{X}$ has the density function

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\Big(-\tfrac{1}{2}(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)\Big)$$

for $\mathbf{x} \in \mathbb{R}^n$.

A plot of the two-dimensional Gaussian probability density function is shown below:



Definition (Marginal distribution)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathbf{X} = (X_1, \ldots, X_n)^T$ a random vector. The distribution of a component X_k, $k = 1, \ldots, n$, is called* ***marginal distribution****. The corresponding distribution function for $z \in \mathbb{R}$ is*

$$F_{X_k}(z) = P(X_k \leq z) = P\big(\{\omega \in \Omega : \ X_k(\omega) \leq z\}\big),$$

and the density function

$$f_{X_k}(z) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{X_1,\ldots,X_n}(x_1, \ldots, x_{k-1}, z, x_{k+1}, \ldots, x_n)\ dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_n.$$

Example: The marginal distribution of a 2–dimensional Gaussian is a Gaussian.

4.3 Expectation

Definition (Expectation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, X a random variable.*
If $X : \Omega \to \{x_i\}$ is a discrete random variable with probability mass function f_X, then

$$\mu := E(X) := \sum_i x_i f_X(x_i)$$

is called the ***expectation*** *of X (or expected value, or mean), if $\sum_i |x_i| f_X(x_i) < \infty$.*
If $X : \Omega \to \mathbb{R}$ is a continuous random variable with probability density function f_X, then

$$\mu := E(X) := \int_{\mathbb{R}} x f_X(x)\ dx$$

is called the ***expectation*** *of X (or expected value, or mean), if $\int_{\mathbb{R}} |x| f_X(x)\ dx < \infty$.*

For measurable $g : \mathbb{R} \to \mathbb{R}$ we define for discrete X

$$E\big(g(X)\big) := \sum_i g(x_i) f_X(x_i)$$

and for continuous X

$$E\big(g(X)\big) := \int_{\mathbb{R}} g(x) f_X(x)\ dx,$$

assuming the sum/integral is well defined.

Properties: Let $(\Omega, \mathcal{F}, P)$ be a probability space, X a random variable.

- If $X \geq 0$, then $E(X) \geq 0$.
- If $X = \chi_A$ for $A \in \mathcal{F}$, then $E(X) = P(A)$.
- Let $X_1, \ldots, X_n$ be random variables with an expectation and $a, b, a_1, \ldots, a_n \in \mathbb{R}$, then

$$\begin{aligned} E(a) &= a \\ E(a + bX) &= a + bE(X) \\ E(a_1X_1 + \ldots + a_nX_n) &= a_1E(X_1) + \ldots + a_nE(X_n). \end{aligned}$$

Definition (Variance, standard deviation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let X be a random variable with expectation $\mu = E(X)$. Then*

$$\sigma^2 := \mathit{Var}(X) := E\big((X - \mu)^2\big)$$

*is called the **variance** of X, and*

$$\sigma := \sqrt{\mathit{Var}(X)}$$

*is called the **standard deviation** of X.*

Properties: Let $(\Omega, \mathcal{F}, P)$ be a probability space, X a random variable.

- If X is discrete we have

$$\mathrm{Var}(X) = \sum_i (x_i - \mu)^2 f_X(x_i),$$

  if X is continuous we have

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x - \mu)^2 f_X(x)\ dx.$$

- $\mathrm{Var}(X) = E(X^2) - E(X)^2$.
- For $a, b \in \mathbb{R}$ we have

$$\begin{aligned} \mathrm{Var}(X) &\geq 0 \\ \mathrm{Var}(a) &= 0 \\ \mathrm{Var}(a + bX) &= b^2\mathrm{Var}(X) \end{aligned}$$

Examples: Let $(\Omega, \mathcal{F}, P)$ be a probability space, let X be a random variable.

- If X is binomially distributed, $X \sim B(n,p)$, that is $P(A) = p$ (where A from the Bernoulli experiment) and $f_X(x_i) = \binom{n}{i} p^i (1-p)^{n-i}$, then

$$E(X) = np, \qquad \mathrm{Var}(X) = np(1-p).$$

- If X is Poisson distributed, $X \sim Poi(\lambda)$, that is $f_X(x_i) = e^{-\lambda} \frac{\lambda^i}{i!}$, then

$$E(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda.$$

- If X is uniformly distributed, $X \sim U(a,b)$, that is $f_X(x) = \frac{1}{b-a} \chi_{(a,b)}(x)$, then

$$E(X) = \tfrac{1}{2}(a+b), \qquad \mathrm{Var}(X) = \tfrac{1}{12}(b-a)^2.$$

- If X is Gaussian distributed, $X \sim \mathcal{N}(\mu, \sigma^2)$, that is $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, then

$$E(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2.$$

Notation: Let $(\Omega, \mathcal{F}, P)$ be a probability space, X random variable with cumulative distribution function F_X and probability mass function / probability density function f_X. To simplify notation (and for other reasons) one introduces the Riemann–Stieltjes integral, or the Lebesgue integral. In our case this allows to write for measurable $g : \mathbb{R} \to \mathbb{R}$

$$\int_{\mathbb{R}} g(x)\ dF_X(x),$$

which is another notation for $\sum_i g(x_i) f_X(x_i)$ if X is discrete, and for $\int_{\mathbb{R}} g(x) f_X(x)\ dx$ if X is continuous.

Again we can generalize these concepts to n dimensions:

Definition (n–dimensional expectation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a random vector with $\mu_k := E(X_k)$. Then the n–**dimensional expectation** of $\mathbf{X}$ is defined as*

$$E(\mathbf{X}) := (\mu_1, \ldots, \mu_n)^T.$$

Theorem. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a random vector with distribution functions $F_{\mathbf{X}}, F_{X_1}, \ldots, F_{X_n}$. If $E(X_k)$ exists for all $k = 1, \ldots, n$, then we have for $\mathbf{x} = (x_1, \ldots, x_n)^T$*

$$E(X_k) = \int_{\mathbb{R}^n} x_k\ dF_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}} x_k\ dF_{X_k}(x_k).$$

Definition (Covariance)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathbf{X} = (X_1, \ldots, X_n)^T$ a random vector where all second moments $E(X_k X_l)$ exist for $k, l \in \{1, \ldots, n\}$. Let $\mu_k := E(X_k)$, then we call for $k, l \in \{1, \ldots, n\}$*

$$\sigma_{kl} := Cov(X_k, X_l) := E\big((X_k - \mu_k)(X_l - \mu_l)\big)$$

*the **covariances** of X_k and X_l. The matrix*

$$\mathbf{C} := \boldsymbol{Cov}(\mathbf{X}) := (\sigma_{kl})_{k,l} \in \mathbb{R}^{n\times n}$$

*is called **covariance matrix** of $\mathbf{X}$.*

*X_k and X_l are called **uncorrelated** if $Cov(X_k, X_l) = 0$.*

Properties: Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a random vector with expectation $\mu = E(\mathbf{X})$ and covariance matrix $\mathbf{C}$.

- $\mathbf{C}$ is positive semi–definite.
- For $k \in \{1, \ldots, n\}$ we have $\mathrm{Cov}(X_k, X_k) = \mathrm{Var}(X_k)$.
- $\mathbf{C} = E\big((\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\big)$.

Definition (Independent random variables)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a random vector with cumulative distribution functions $F_{\mathbf{X}}$ and $F_{X_1}, \ldots, F_{X_n}$. The components X_k of $\mathbf{X}$ are called **independent** if*

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n).$$

This can be formulated equivalently using the density functions.

Definition (iid)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let X and $X_1, \ldots, X_n$ be random variables. $X_1, \ldots, X_n$ are called **independent identically distributed** or in short **iid** excactly if the X_k are independent and $X \sim X_k$ for all $k = 1, \ldots, n$.*

Definition (Correlation coefficient)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathbf{X} = (Z, Y)^T$ a random vector with covariance matrix $\mathbf{C} = Cov(\mathbf{X})$. If $Var(Z) > 0$ and $Var(Y) > 0$, then*

$$\rho(Z, Y) := \frac{Cov(Z, Y)}{\sqrt{Var(Z)\, Var(Y)}}$$

*is called the **correlation coefficient** of Z and Y.*

Properties: If Z and Y are uncorrelated, then $\rho(Z, Y) = 0$. The correlation coefficient is a measure of the linear relationship between Z and Y, as we have $\rho(Z, Y) \in [-1, 1]$ and $\rho(Z, Y) = \pm 1$ exactly if $P(Y = a + bZ) = 1$ for $a, b \in \mathbb{R}$, $b \neq 0$.

Example: Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $(Z, Y)^T \sim \mathcal{N}(\mu, \mathbf{C})$, i.e. a 2–dimensional Gaussian random variable. Assume $\mu = (\mu_z, \mu_y)^T$, choose $\sigma_z, \sigma_y, \rho \in \mathbb{R}$ with $\sigma_z, \sigma_y > 0$, $\rho \in (-1, 1)$ such that

$$\mathbf{C} := \mathbf{Cov}(Z, Y) = \begin{pmatrix} \sigma_z^2 & \rho\sigma_z\sigma_y \\ \rho\sigma_z\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Then we have $E(Z) = \mu_z$, $E(Y) = \mu_y$, $\mathrm{Var}(Z) = \sigma_z^2$, $\mathrm{Var}(Y) = \sigma_y^2$ and $\rho(Z, Y) = \rho$.

Theorem. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, (Z, Y) random variables with expectation $\mu = (\mu_z, \mu_y)$ and covariance matrix $\mathbf{C}$. If Z and Y are independent, then they are uncorrelated. The opposite implication is false.*

If $(Z, Y) \sim \mathcal{N}(\mu, \mathbf{C})$, then: (Z, Y) uncorrelated $\iff$ (Z, Y) independent.

4.4 Conditional expectation

Sometimes it is helpful to use the concept of *conditioning* to simplify a problem.

Definition (Conditional distribution and expectation given an event)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : \Omega \to \{x_i\}$ a discrete random variable. Let $B \in \mathcal{F}$ with $P(B) > 0$ and $A_i = \{\omega \in \Omega : \ X(\omega) = x_i\}$, then*

$$f(x_i|B) := P(X = x_i|B) = P(A_i|B) = \frac{P(A_i \cap B)}{P(B)}$$

*defines a **distribution of** X **under condition** B. We call*

$$E(X|B) := \sum_i x_i \ f(x_i|B)$$

*the **conditional expectation of** X **given** B.*

Definition (Conditional distribution)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, (X, Y) a random vector with joint probability density function $f_{X,Y}$ and marginal distribution f_Y. Then the **conditional distribution** $f_{X|Y}$ **of** X **given** Y is defined as*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \qquad \text{for all } y \in \mathbb{R} \text{ with } f_Y(y) > 0.$$

$f_{X|Y}$ is again a probability density function, the corresponding random variable is:

Definition (Conditional expectation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, (X, Y) a random vector and $f_{X|Y}$ the conditional distribution of X given Y. Using*

$$E(X|Y) := E(X|Y = y) := \int_{\mathbb{R}} x f_{X|Y}(x|y) \ dx$$

*we define a new random variable $Z : \Omega \to \mathbb{R}$ by $Z := E(X|Y)$, which is called the **conditional expectation of** X **given** Y.*

Example: roll a die until you get a 6. Define random variables $Y =$ total number of rolls, $X =$ number of 1s rolled. If you now compute $E(X|Y = y)$, this means $y - 1$ rolls were not a 6, and the y-th roll was 6. Give this, we have $X \sim B(n, p)$ with $n = y - 1$ and $p = \frac{1}{5}$, which implies

$$E(X|Y = y) = np = \frac{1}{5}(y - 1).$$

If you now perform an experiment, you get a series of $1, \ldots, 5$ and then a 6, you can encode this series as ω. Compute now $y = Y(\omega)$, i.e. the total number of rolls and then compute $E(X|Y = y)$, then $\omega \mapsto E(X|Y = y)$ defines a random variable Z, here we have

$$Z = E(X|Y) = \frac{Y - 1}{5}.$$

Properties: Let $(\Omega, \mathcal{F}, P)$ be a probability space, (X, Y, Z) a random vector.

- We have $E\big(E(X|Y)\big) = E(X)$.
- For $a, b \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ measurable we have

$$\begin{aligned} E(a|Y) &= a \\ E(aX + bZ|Y) &= aE(X|Y) + bE(Z|Y) \\ E(X|Y) &= E(X) \quad \text{if } X, Y \text{ independent} \\ E(Xg(Y)|Y) &= g(Y)E(X|Y) \end{aligned}$$

4.5 Estimators

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Assume we have a set of random variables $X_1, \ldots, X_n$ that are iid. If $X_i \sim Poi(\lambda_i)$, we might not know λ_i and would have to estimate it from the $X_1, \ldots, X_n$. In general we say we have a vector of unknown parameters $\theta = (\theta_1, \ldots, \theta_m)^T$.

Definition (Estimator)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X_1, \ldots, X_n$ iid random variables. To estimate a set of parameters $\theta \in \Theta \subset \mathbb{R}^m$ we define a measurable function*

$$T : \mathbb{R}^n \to \Theta, \qquad T(X_1, \ldots, X_n) = \hat{\theta} \in \Theta.$$

*This function is called **estimator** and is itself a random variable on a different probability space.*

*An estimator T is called **unbiased** if*

$$E\big(T(X_1, \ldots, X_n)\big) = \theta.$$

*The difference $E\big(T(X_1, \ldots, X_n)\big) - \theta$ is called **bias** of the estimator T.*

There are also *consistent* estimators, however the definition requires convergence in probability, which is out of the scope of this lecture.

Now for iid random variables $X_1, \ldots, X_n$ with unknown parameters $\theta \in \mathbb{R}^m$ and density $f(x; \theta)$ (we add θ here to indicate that we don't know it), the joint n–dimensional density of the random vector $(X_1, \ldots, X_n)$ is

$$l(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

which is called the **likelihood function**. The so–called **maximum likelihood method** is then to find an estimate θ_{ML} for the unknown θ such that for the occurence $(x_1, \ldots, x_n)$ we have maximum likelihood, i.e.

$$l(x_1, \ldots, x_n; \theta_{ML}) \geq l(x_1, \ldots, x_n; \theta) \qquad \forall \theta \in \mathbb{R}^m.$$

Alternatively we can also study the **log–likelihood function**

$$L(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

to ease computations.

Definition (ML estimator)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X_1, \ldots, X_n$ iid random variables with density $f(x;\theta)$ with unknown parameters $\theta \in \mathbb{R}^m$. Then every solution θ_{ML} to*

$$\theta_{ML} = argmax\big\{L(x_1, \ldots, x_n; \theta) : \ \theta \in \mathbb{R}^m\big\} = argmax\big\{l(x_1, \ldots, x_n; \theta) : \ \theta \in \mathbb{R}^m\big\}$$

*is called **ML estimator** or **maximum likelihood estimator** for θ. θ_{ML} may not exist or may not be unique.*

If the X_k are Gaussian random variables, the ML estimator corresponds to the least squares solution.

4.6 Expectation maximization

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_l)^T$ be iid random vectors with parameters $\theta \in \mathbb{R}^m$ and density functions $f(\cdot;\theta)$. Here we will call the random vector $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})^T$ our complete data, which is split into observed data $\mathbf{Y}$ and unobserved data $\mathbf{X}$ (also called hidden parameters). Using the ML estimator we would form the log–likelihood function

$$L(\mathbf{z};\theta) = \sum_i \ln f(z_i;\theta)$$

with $\mathbf{z} = (z_1, \ldots, z_{n+l})^T$. As we don't know the $\mathbf{X}$ component of $\mathbf{Z}$, computing an ML estimate can be problematic.

Instead, we use the tool of conditional expectation to simplify our problem by computing instead

$$E\big(L(\mathbf{z};\theta)|\mathbf{y};\theta_{current}\big) \qquad \textbf{(E-step)}$$

and computing the ML estimator on that conditional expectation

$$\theta_{ML} = \text{argmax}_\theta \ E\big(L(\mathbf{z};\theta)|\mathbf{y};\theta_{current}\big) \qquad \textbf{(M-step)}$$

Starting with an initial guess θ^0 and iteratively computing

$$\theta^{k+1} = \text{argmax}_\theta \ E\big(L(\mathbf{z};\theta)|\mathbf{y};\theta^k\big)$$

is called the **expectation maximization method**.

While this procedure guarantees $L(\mathbf{z},\theta^{k+1}) \geq L(\mathbf{z},\theta^k)$, convergence is not guaranteed in the general case.

4.7 EM in emission tomography

In emission tomography we want to recover a function $f : V \to \mathbb{R}$, that maps a volume of interest $V \subset \mathbb{R}^n$ to the average number of radioactive decays that happened during the measurement period. Like previously, we discretize f using a finite set of basis functions $b_i : V \to \mathbb{R}$, $i \in I$, $|I| = n$, such that f can be approximated by

$$\hat{f} = \sum_{i \in I} x_i b_i,$$

or in other words: we represent $\hat{f}$ by a coefficient vector $(x_i)_{i\in I}$. The task is then to "reconstruct" $\hat{f}$ from a set of measurements $(m_j)_{j\in J}$ acquired by detectors $j \in J$.

Contrary to the series expansion method covered in section 1.7, we now put this into a probabilistic framework.

We denote

$$a_{ji} := P(\text{emission detected in detector } j \mid \text{emission occured in basis } i)$$

and assume we know a way to compute this (for example using a model or simulations). Then we look at the iid random vector $\mathbf{Z} = (X_{ji})_{j\in J, i\in I}$, where each X_{ji} is a Poisson distributed random variable, $X_{ji} \sim Poi(\lambda_{x_{ji}})$, denoting the number of emissions in basis i detected by detector j. Then

$$\lambda_{x_{ji}} = E(X_{ji}) = x_i a_{ji}. \tag{15}$$

Our measurements m_j, $j \in J$, correspond now to the random variables $Y_j = \sum_{i\in I} X_{ji}$ with $Y_j \sim Poi(\lambda_{y_j})$, where

$$\lambda_{y_j} = E(Y_j) = E\left(\sum_{i\in I} X_{ji}\right) = \sum_{i\in I} x_i a_{ji} = m_j.$$

Instead of treating this as a linear system and solving it (like we did before), we now try to compute an ML estimate instead.

The parameters we want to estimate are the $\theta := (x_i)_{i\in I}$. The likelihood function to consider here is then

$$l(\mathbf{z};\theta) = \prod_{j\in J}\prod_{i\in I} f(z_{ji};\theta) = \prod_{j\in J}\prod_{i\in I} e^{-\lambda_{x_{ji}}} \frac{\lambda_{x_{ji}}^{z_{ji}}}{z_{ji}!},$$

or the log–likelihood

$$\begin{aligned} L(\mathbf{z};\theta) &= \sum_{j\in J}\sum_{i\in I} -\lambda_{x_{ji}} + z_{ji}\ln\lambda_{x_{ji}} - \ln z_{ji}! \\ &= \sum_{j\in J}\sum_{i\in I} -E(X_{ji}) + z_{ji}\ln E(X_{ji}) - \ln z_{ji}! \\ &= \sum_{j\in J}\sum_{i\in I} -x_i a_{ji} + z_{ji}\ln(x_i a_{ji}) - \ln z_{ji}! \end{aligned}$$

using (15). This log–likelihood can be shown to be concave.

Now we employ the EM method from the previous section,

$$\theta^{k+1} = \operatorname{argmax}_\theta \, E\big(L(\mathbf{z};\theta) \mid \mathbf{Y} = \mathbf{m};\theta^k\big),$$

where $\mathbf{m} = (m_j)_{j\in J}$ is a realization of the random vector $\mathbf{Y} = (Y_j) = (\sum_{i\in I} X_{ji})$.

Step 1: Expectation. We have to compute the conditional expectation

$$\begin{aligned} E\big(L(\mathbf{z};\theta) \mid \mathbf{m};\theta^k\big) &= E\left(\sum_{j\in J}\sum_{i\in I} -x_i a_{ji} + z_{ji}\ln(x_i a_{ji}) - \ln z_{ji}! \;\middle|\; \mathbf{m},\theta^k\right) \\ &= \sum_{j\in J}\sum_{i\in I}\Big(- x_i a_{ji} + E\big(X_{ji} = z_{ji}|\mathbf{m},\theta^k\big)\ln(x_i a_{ji}) - E\big(X_{ji} = \ln z_{ji}!|\mathbf{m},\theta^k\big)\Big). \end{aligned}$$

The last term will vanish in Step 2, but we need to compute $E\big(X_{ji} = z_{ji}|\mathbf{m}, \theta^k\big)$. Fortunately we know (see exercises), that X_{ji} given $\sum_{i\in I} X_{ji}$ is binomially distributed $B(n,p)$ with $n = m_j$ and $p = \frac{E(X_{ji})}{\sum_{l\in I} E(X_{jl})}$. With equation (15) and $\theta^k = (x_i^k)_{i\in I}$ that yields

$$E\big(X_{ji} = z_{ji}|\mathbf{m}, \theta^k\big) = m_j \ \frac{x_i^k a_{ji}}{\sum_{l\in I} x_l^k a_{jl}}.$$

Step 2: Maximization. As usual we set the derivate to zero,

$$\frac{d}{d\theta}\ E\big(L(\mathbf{z};\theta) \mid \mathbf{m};\theta^k\big) \stackrel{!}{=} 0.$$

We do this component–wise, $l \in \{1,\dots,n\}$:

$$\frac{\partial}{\partial x_l}\ E\big(L(\mathbf{z};\theta) \mid \mathbf{m};\theta^k\big) = \sum_{j\in J} -a_{jl} + \sum_{j\in J} \frac{1}{x_l} E\big(X_{jl} = z_{jl}|\mathbf{m}, \theta^k\big) \stackrel{!}{=} 0$$

which is equivalent to

$$x_l = \frac{\sum_{j\in J} E\big(X_{jl} = z_{jl}|\mathbf{m}, \theta^k\big)}{\sum_{j\in J} a_{jl}} = \frac{x_l^k}{\sum_{j\in J} a_{jl}} \sum_{j\in J} \frac{m_j a_{jl}}{\sum_{i\in I} x_i^k a_{ji}}.$$

This is finally known as the so–called **MLEM algorithm**: using a starting estimate $\theta^0 = (x_1^0,\dots,x_n^0) \geq 0$ we iterate

$$x_l^{k+1} = \frac{x_l^k}{\sum_{j\in J} a_{jl}} \ \sum_{j\in J} a_{jl} \frac{m_j}{\sum_{i\in I} x_i^k a_{ji}} \qquad \forall l \in \{1,\dots,n\}.$$

Convergence of MLEM to a ML estimator can be proven (see literature), however uniqueness cannot be guaranteed. A nice feature of MLEM is

$$\sum_{i\in I} x_i^k = \sum_{j\in J} m_j,$$

that is the total number of emissions counted is preserved, and furthermore $x_i^k \geq 0$.

As a closing remark, one can show that MLEM is actually an iterative projection algorithm like ART, using a different concept of orthogonal projections (entropy-based). MLEM can also be interpreted as a gradient descent algorithm,

$$x_l^{k+1} = x_l^k + \frac{x_l^k}{\sum_{j\in J} a_{jl}}\ \frac{\partial}{\partial x_l} L(\mathbf{z};\theta^k)$$